Andrew Nee
andrewnee29@gmail.com

Predicting MLB Pitcher Roles

Background

Traditionally, baseball pitchers have been split up into two roles: starting pitcher(SP), relief pitcher(RP). A starting pitcher starts in the beginning of the game and is expected to pitch the majority of the game, around five innings. Starters usually have many different pitches and pitching sequences as they will have to face batters multiple times. A relief pitcher is any pitcher that follows a starting pitcher and is expected to throw anywhere from one to four innings. Relievers usually have around a couple of pitches that they rely on as they only have to face batters one or two times. Starters run a marathon while relievers run a sprint.

However, in 2023, pitchers threw an average of 5.1 innings per start, a stark decrease from 5.9 in 2013. Subsequently, MLB teams averaged 526 relief appearances in 2023 compared to 478 in 2013. Multi-inning relief appearances also increased from 102 to 131 in the same timespan. These numbers reflect change in how modern MLB pitching staffs see pitching roles. First, they are becoming increasingly amorphous in nature. Second, pitching roles are more fluid than ever. Third, pitchers are better utilized across specific and multiple innings where the game situation enables teams to get the most value out of the pitcher. This poses a problem: pitchers are currently serving in a role that is suboptimal for their performance and value.

Dataset Introduction

The dataset, “fangraphs_season_level.csv”, was collected from fangraphs.com’s library. It is composed of every statistic that fangraph utilizes, both advanced and standard. Fangraph collects this data utilizing Statcast, a high-accuracy camera tracking system that collects and

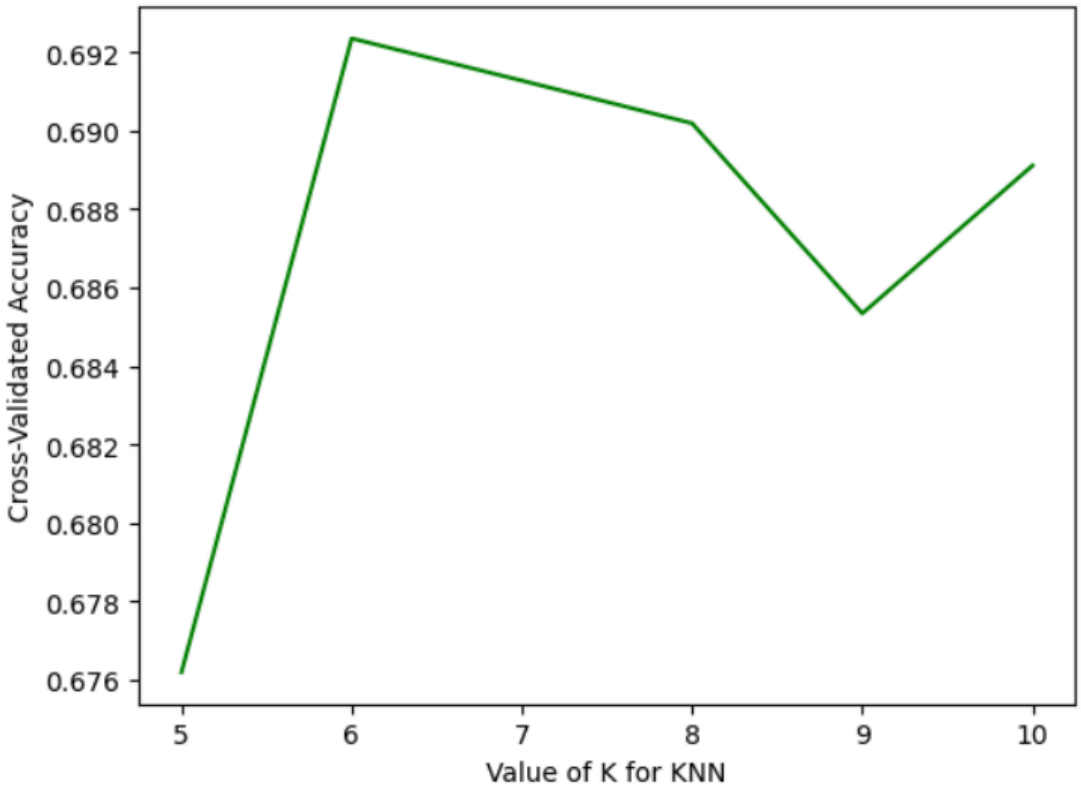
analyzes all on-field movement. MLB has installed these cameras at all major ballparks and it is well-known and public data, so there are minimal privacy and ethical concerns about the collection. As for bias, the only human element in the datasets are called balls and strikes. These are called by the home plate umpire. While they are usually consistent, there may be some variation and inconsistency in their called balls and strikes, which may skew the data, but is negligible as the data includes games with many different home plate umpires over the 2021-2023 seasons.

The statistics I believe to be the most relevant a pitcher's qualities are FIP, xFIP, SIERA, Stuff+, Location+, Pitching+, WAR, WPA/LI, K%, BB%, Soft%, Med%, Hard%, GB%, LD%, FB%, and SwStr%. FIP, xFIP, and SIERA are all similar to a well known stat, ERA (earned run average), which is how many runs a pitcher allows in 9 innings. FIP only uses outcomes that do not include defense in order to single out a pitcher's performance. xFIP not only excludes these outcomes, but also uses league average home run(HR)/fly ball(FB) rate as it can be variable. Both of these stats would tend to underrate pitchers that can generate fly balls(FB) while also limiting home runs. SIERA is similar to xFIP, but gives more credit to strikeouts(K), and less blame for FBs. This is due to pitchers with high K% tend to have lower batting averages for balls in play(BABIP) and HR/FB%. Stuff+ quantifies how "nasty" a pitch is with its physical characteristics of the ball such as release point, velocity, vertical and horizontal movement, spin rate, etc. Location+ is a count and pitch type-adjusted judge of a pitcher's command. This is due to the fact that an effective breaking ball in a 2-0 (balls/strikes) count and a 0-2 count would differ. Pitching+ is an overall look on a pitcher's process and uses aspects of Stuff+ and Location+; it has been shown to be very predictively powerful in relation to other advanced pitching stats. Wins above replacement(WAR) measures how many more wins a player would

generate you compared to league average. Win probability average(WPA) is a change in win expectancy from one plate appearance to another. However, it doesn't necessarily tell you if a pitcher pitched well, but if the team had a positive outcome when they were on the field. Leverage index(LI) attempts to quantify how "on the line" the game is at a particular moment. By dividing both of these statistics into WPA/LI, it shows how much a player provided regardless of the state of the game. BB% is walk percentage. Soft%, Med%, Hard% measures the quality of contact on batted balls because generally, the harder a ball is hit, the more likely it will fall for a hit. GB% and LD% are ground ball percentage and line drive percentage, respectively. Finally, Swung Strikes(SwStr) is the total number of swings and misses/total pitches.

Data Science Approaches

Because I am attempting to predict pitchers into a SP and RP role, I chose to use a K nearest neighbors(KNN) classifier, which will use pitchers' Euclidean proximity to one another to make these predictions. The 2021-2023 seasons will be weighted linearly more towards the 2023 season as it is important to weigh the more current seasons. I then cross validate the model to find the most accurate number of neighbors to determine the role classification.



Lastly, I chose to consider pitchers with both starting and relief appearances as different as their specific role and game condition variate.

	PlayerId	NameASCII	Role	Preds	K_pct	BB_pct	FIP	xFIP	SIERA	WAR	...	SwStr_pct
0	18	Neftali Feliz	0	0	0.1579	0.0526	6.420031	5.375481	4.519107	-0.174990	...	0.104500
1	1157	Tommy Hunter	0	1	0.2044	0.0641	4.883428	4.436076	3.934551	-0.141653	...	0.088383
2	1157	Tommy Hunter	1	0	0.1429	0.0000	2.170030	3.934390	4.262984	0.067579	...	0.000000
3	1159	Andrew Romine	0	0	0.2000	0.0000	14.170030	4.698750	3.962118	-0.051028	...	0.133300
4	1246	Matt Bush	0	0	0.2548	0.0962	7.887948	4.404207	3.929122	-0.362258	...	0.128283
...	...	...	...	...	...	...	...	...	...	...	...	...
5104	27758	Kyle Harrison	1	0	0.2381	0.0748	5.533884	5.013914	4.445423	-0.112184	...	0.093500
5105	29832	Jackson Wolf	1	0	0.0455	0.0455	3.455040	5.108393	5.288328	0.105420	...	0.026300
5106	29911	Andrew Abbott	1	0	0.2614	0.0959	4.197116	4.562998	4.334174	2.156510	...	0.109100
5107	29928	Dylan Dodd	1	1	0.0915	0.0732	6.924946	6.502845	6.185948	-0.504884	...	0.084500
5108	30094	Jordan Wicks	1	1	0.1633	0.0748	4.697346	4.300826	4.828501	0.279402	...	0.099500

Here are the average stats of a starting pitcher vs. relief pitcher.

0	K_pct	0.202051
1	BB_pct	0.093717
2	FIP	5.635269
3	xFIP	4.962899
4	SIERA	4.872335
5	WAR	0.552526
6	WPA	-0.090868
7	WPA_to_LI	-0.120912
8	SwStr_pct	0.101466
9	Soft_pct	0.160237
10	Med_pct	0.497923
11	Hard_pct	0.341836
12	GB_pct	0.410526
13	LD_pct	0.201085
14	FB_pct	0.388389
15	Stuff_plus	95.130869
16	Location_plus	98.757181
17	Pitching_plus	97.558659

0	K_pct	0.197518
1	BB_pct	0.101735
2	FIP	5.476503
3	xFIP	5.156392
4	SIERA	4.711119
5	WAR	0.090038
6	WPA	0.028987
7	WPA_to_LI	0.004271
8	SwStr_pct	0.100426
9	Soft_pct	0.156743
10	Med_pct	0.514558
11	Hard_pct	0.328699
12	GB_pct	0.417982
13	LD_pct	0.197168
14	FB_pct	0.384850
15	Stuff_plus	89.606842
16	Location_plus	95.752061
17	Pitching_plus	94.631173

Results and Conclusions

Here are the actual and predicted roles of the top 5 voted of the 2023 AL Cy Young and NL Cy Young awards.

	PlayerId	NameASCII	Role	Preds
494	17995	Logan Webb	0	0
495	17995	Logan Webb	1	1
770	27498	Spencer Strider	0	0
771	27498	Spencer Strider	1	1
1359	20638	Luis Castillo	0	0
3191	12768	Sonny Gray	1	1
3203	13125	Gerrit Cole	1	0
3239	14107	Kevin Gausman	1	1
3289	15689	Luis Castillo	1	1
3421	19291	Zac Gallen	1	1
3994	24586	Kyle Bradish	1	0
4584	13543	Blake Snell	1	0

It correctly predicts seven out of the ten pitchers. While Gerrit Coles, Kyle Bradish, and Blake Snell's stats may better reflect a relief pitcher. All of these pitchers have either surpassed or have closely tied their total innings pitched in 2023. I have only included one value added stat, WAR, which takes into account how many innings and games the pitcher has pitched in. My initial logic was the amount a pitcher pitches does not really correlate to their pitch quality. That being said, all three of these pitchers have shown tremendous recovery, health, and consistency throughout the season and are obviously extremely effective at the starting pitcher role. These pitchers have rarely relief pitched in their career, in any, so pitching staffs should consider them

for relieving in high-intensity game situations if they feel recovered from their starts, especially in the playoffs.

Future Work

Some things to consider for potential improvements on this project is season weight, potential for a third or more classifications of pitchers that do not fit well in a starting or relieving role. The season weight I used was arbitrary, I just focused on weighing the more current seasons. An exponential weight may be more effective at predicting a current pitcher's abilities. Additional pitcher classifications could include specialists such as left-handed specialists or batter specific specialists, and long relievers, middle relievers, and closers. While these roles may already exist within MLB pitching staffs, more work would need to be done to quantify such roles for classification.