Dylan Guzman

Jan. 15, 2025

Project Outline

The goal of this project is to find a range of minor league player statistics that have a high percentage of players who made it to the major league level. The data used for this project was obtained from fangraphs.com, and includes statistics from every player at the A level from 2017 to 2019 that have at least 100 plate appearances. If you are unfamiliar with the minor league system, the A level is a low level of the minor leagues, as players must get promoted to High-A, AA, and AAA, before they are called up to the major league level. Very few players get promoted all the way to the major leagues; in fact, only around ten percent of the players from the previously mentioned dataset had a long stint in the major leagues. That being said, the baseline for me was to find a range of player statistics that can predict players to make it to the MLB at a rate higher than ten percent.

Model Process

My first idea was to take a single statistic, line drive rate, and split my data into groupings based on it, and see what percentage of players in each group made it to the MLB. I created a training dataset with roughly 60% of my total data and found the minimum and maximum line drive rate. I found the midpoint between the two points, establishing that any player above the midpoint would be in group A, and any player below would be in group B. I then repeated this process periodically increasing the number of groups all of the way up to twenty. I then found the percentage of players that had at least 500 plate appearances in the MLB for each group.

My next step was to incorporate other statistics besides line drive rate. I created a similar model, except it also created groupings for walk/strikeout rate. It then took every grouping from line drive rate, and paired it with every grouping from walk/strikeout rate, and then evaluated how many players made it to the MLB with statistics in both ranges.

A lot of the statistics I am using are highly correlated with each other. For instance, fly ball rate and ground ball rate have an r-value of -.8637. Moving forward, I decided to divide the statistics into four different categories to avoid multicollinearity: batted ball stats, batted direction stats, plate discipline stats, and hitting results. The groupings can be seen below.

Batted Ball	Batted Direction	Plate Discipline	Hitting Results
Line Drive (LD) % Ground Ball (GB) % Fly Ball (FB) % Home Run/FB GB/FB	Pull % Center % Oppo %	Walk (BB) % Strikeout (K) % BB/K	Batting Average On Base % Slugging % Isolated Power On Base + Slugging %

The final step for my project was to create groupings for each individual statistic, then compare each group with every other group for every statistic in the other statistical categories. It took my computer nearly two hours to run this, I had to shorten the amount of groups to ten for each statistic. The results can be seen below, sorted by overall_success, which is the percentage of players within the grouping that had at least 500 plate appearances in the MLB. Success represents the success rate within the training dataset, and test_success is the success rate in the testing set. Upper and Lower represent the upper and lower bounds of their respective metric.

Total and test_total refer to the number of results in each group for the training and testing dataset respectively.

success	total	test_success	test_total	overall_success	metric1	lower	upper	metric2	lower2	upper2	metric3	lower3	upper3	metric4	lower4	upper4
0.3636364	11	0.8000000	5	0.5000000	GB.	0.45100000	0.55400000	Pull.	0.4335000	0.5257500	K._A	0.07400000	0.16250000	AVG.y	0.2735000	0.3287500
0.2727273	11	0.8333333	6	0.4705882	GB.	0.45100000	0.55400000	Oppo.	0.2400000	0.3200000	K._A	0.07400000	0.16250000	AVG.y	0.2735000	0.3287500
0.4545455	11	0.3333333	3	0.4285714	LD.	0.16000000	0.19850000	Oppo.	0.2133333	0.2666667	BB.K_A	0.31666667	0.55333333	OPS_A	0.7470000	0.8550000
0.3571429	14	0.4736842	19	0.4242424	GB.FB	0.43000000	1.51333333	Cent.	0.2266667	0.2973333	K._A	0.07400000	0.19200000	OBP_A	0.3693333	0.4480000
0.2307692	13	0.8333333	6	0.4210526	LD.	0.08300000	0.16000000	Oppo.	0.2666667	0.3733333	K._A	0.07400000	0.19200000	ISO_A	0.1160000	0.2110000
0.4166667	12	0.4210526	19	0.4193548	FB.	0.31600000	0.46200000	Cent.	0.2266667	0.2973333	K._A	0.07400000	0.19200000	OBP_A	0.3693333	0.4480000
0.2500000	12	0.8000000	5	0.4117647	LD.	0.14075000	0.19850000	Pull.	0.4335000	0.5257500	K._A	0.07400000	0.16250000	OPS_A	0.7470000	0.9090000
0.3333333	12	0.5000000	10	0.4090909	GB.FB	0.43000000	1.51333333	Cent.	0.2266667	0.2973333	K._A	0.07400000	0.19200000	AVG.y	0.3103333	0.3840000
0.2727273	11	0.5000000	16	0.4074074	GB.FB	0.43000000	1.51333333	Oppo.	0.2666667	0.3733333	K._A	0.07400000	0.19200000	OBP_A	0.3693333	0.4480000
0.2727273	11	0.7500000	4	0.4000000	GB.	0.42157143	0.48042857	Oppo.	0.2514286	0.2971429	BB._A	0.06914286	0.09671429	ISO_A	0.1024286	0.1431429
0.2727273	11	0.7500000	4	0.4000000	GB.FB	1.24250000	2.05500000	Oppo.	0.2400000	0.3200000	K._A	0.07400000	0.16250000	AVG.y	0.2735000	0.3287500
0.4000000	15	0.4000000	10	0.4000000	GB.	0.45100000	0.65700000	Pull.	0.4335000	0.6180000	K._A	0.07400000	0.25100000	SLG_A	0.4150000	0.6230000
0.4545455	11	0.2500000	4	0.4000000	LD.	0.18200000	0.21500000	Cent.	0.2468571	0.2771429	BB._A	0.06914286	0.09671429	OBP_A	0.3131429	0.3468571
0.4545455	11	0.2500000	4	0.4000000	GB.FB	0.43000000	2.05500000	Pull.	0.4335000	0.6180000	BB._A	0.11050000	0.20700000	AVG.y	0.2735000	0.3840000
0.2666667	15	0.6250000	8	0.3913043	GB.FB	1.24250000	2.05500000	Pull.	0.4335000	0.5257500	BB._A	0.06225000	0.11050000	AVG.y	0.2735000	0.3287500
0.2727273	11	0.5000000	12	0.3913043	GB.	0.38233333	0.51966667	Cent.	0.2266667	0.2973333	K._A	0.07400000	0.19200000	AVG.y	0.3103333	0.3840000
0.4375000	16	0.2857143	7	0.3913043	FB.	0.35771429	0.42028571	Oppo.	0.2514286	0.2971429	K._A	0.22571429	0.27628571	AVG.y	0.2261429	0.2577143
0.3636364	11	0.4285714	7	0.3888889	GB.FB	0.43000000	1.51333333	Cent.	0.2266667	0.2973333	BB._A	0.07833333	0.14266667	SLG_A	0.4843333	0.6230000
0.4166667	12	0.3333333	6	0.3888889	GB.FB	0.43000000	2.05500000	Oppo.	0.1600000	0.3200000	BB._A	0.11050000	0.20700000	AVG.y	0.2735000	0.3840000
0.2727273	11	0.5000000	10	0.3809524	LD.	0.14075000	0.19850000	Pull.	0.4335000	0.5257500	BB._A	0.06225000	0.11050000	AVG.y	0.2735000	0.3287500
0.4545455	11	0.3000000	10	0.3809524	FB.	0.35771429	0.42028571	Pull.	0.4071429	0.4598571	K._A	0.22571429	0.27628571	AVG.y	0.2261429	0.2577143
0.2500000	12	0.7500000	4	0.3750000	GB.FB	1.08000000	1.73000000	Oppo.	0.2240000	0.2880000	BB._A	0.09120000	0.12980000	OBP_A	0.3536000	0.4008000
0.2727273	11	0.6000000	5	0.3750000	HR.FB	0.05466667	0.10933333	Oppo.	0.2133333	0.2666667	BB._A	0.04616667	0.07833333	ISO_A	0.1160000	0.1635000
0.2857143	14	0.4444444	18	0.3750000	GB.	0.38233333	0.51966667	Cent.	0.2266667	0.2973333	K._A	0.07400000	0.19200000	OBP_A	0.3693333	0.4480000
0.3636364	11	0.4000000	5	0.3750000	FB.	0.43280000	0.52040000	Oppo.	0.2240000	0.2880000	BB._A	0.05260000	0.09120000	OBP_A	0.3064000	0.3536000
0.2105263	19	0.7500000	8	0.3703704	FB.	0.27950000	0.38900000	Pull.	0.4335000	0.5257500	BB._A	0.06225000	0.11050000	SLG_A	0.4150000	0.5190000

The sixth row appears to be the best, as it has the highest success rate while maintaining a negligible difference in success rates among the testing and training datasets. There are currently 13 players at the A-level with statistics that fall in that range, none of which are regarded as top 100 prospects according to mlb.com. Perhaps this project can act as a way to find under the radar prospects that MLB analysts are overlooking.