**Week 4 Report: Original Formula Predictions**

Jai Arora, Robert Banas, Bailey Harrison, Simran Mallik

Variables Selected

**Hitters:** As a team, we thought that On-base Plus Slugging (OPS), Weighted Runs Created + (wRC+), and Home Run to Fly Ball Rate (HR/FB) were the best metrics used to evaluate a hitters potential to make it to the MLB. We wanted to use OPS because it combines a hitters on-base percentage and slugging percentage. Meaning how well a hitter can reach base, with how well he can hit for average and for power. Next, we used wRC+ as our second metric to evaluate hitters. wRC+ takes the statistic Runs Created and adjusts that number to account for important external factors, like ballpark or era. It is useful because wRC+ quantifies run creation and normalizes it. Then, our last metric we used to evaluate hitters was HR/FB. Home Run to Fly Ball rate is the ratio of home runs a player hits out of their total number of fly balls. A hitters HR/FB ratio can be useful in providing context about how sustainable their power is. In our formula, we weighted OPS (40%) and wRC+ (40%) more than HR/FB (20%) because we felt like those metrics were more important. To be a good baseball player, you do not need to hit a lot of homeruns. For example, you can be a high batting average hitter with speed and have a low HR/FB and still be very good.

**Pitchers:** As a team, we thought that Expected Fielding Independent Pitching (xFIP), Walks and Hits Per Inning Pitched (WHIP), and Home Run Per Nine Innings (HR/9) were the best metrics to evaluate a pitchers potential to make it to the MLB. We used xFIP because it is a better version of ERA and FIP. It shows whether the pitcher performs well at limiting runs scored, independent of his team’s poor or positive fielding. We chose xFIP over regular FIP as it, “replaces a pitcher’s home run total with an estimate of how many home runs they should have allowed given the number of fly balls they surrendered while assuming a league average home run to fly ball percentage.” Next, we used WHIP as our second metric. WHIP shows how well a pitcher has kept runners off the basepaths. The formula is the sum of a pitcher's walks and hits, divided by his total innings pitched. Then, our last metric we used was HR/9. Home Run Per Nine Innings represents the average number of home runs allowed by a pitcher on a nine-inning scale. HR/9 is a statistic that is in the control of the pitcher, because defensive positioning plays no factor. In our formula we weighted xFIP (40%) and WHIP (40%) more than HR/9 because we felt like those metrics were more important. Certain pitchers can have success with a high HR/9 rate, as long as they manage to limit their baserunners otherwise, leading to fewer multi-run home runs. However, some pitchers have mastered the art of allowing fly balls but not many home runs.

Data Cleaning

The data that was analyzed was pulled from FanGraphs. There were a total of six MiLB datasets that were cleaned and analyzed (2019 A+ Hitters, 2019 A+ Pitchers, 2019 AA Hitters, 2019 AA Pitchers, 2019 AAA Hitters, 2019 AAA Pitchers) in Excel. The metrics analyzed for all the hitters included On-base Plus Slugging (OPS), Weighted Runs Created + (wRC+), and Home Run to Fly Ball Rate (HR/FB). The metrics analyzed for all the pitchers included Home Run Per Nine Innings (HR/9), Expected Fielding Independent Pitching (xFIP), and Walks and Hits Per Inning Pitched (WHIP). The advanced and batted dataset were used for hitters and the advanced dataset was used for pitchers. *Fangraphs* included an automatic filter that only displayed the statistics for players who were ‘qualified’ based off of the minimum Plate Appearance and Innings Pitched (PA/IP). Additionally, the dataset was filtered for players who were between the ages of 18-25. Players who were 16-17 years old would be more expensive to maintain and require more equipment since they would remain in the MiLB for a longer period of time than older players. The younger player prospects also tend to be international players, which means that the expense is even larger. Players aged 25 in the MiLB were no longer deemed as a viable prospect and were unlikely to reach Major League Baseball (MLB) player status, and therefore were not included in our cleaned dataset.

For each dataset, the total score of each player was calculated using a revised formula which includes multiplying weights by a player’s relative ranking for each metric. This is explained in depth in the “Changes in Our Formula Section”. Once the total score was calculated for all players in each dataset, the players were ranked by total score. Then, we took the top 19 people in each dataset because the 2019 AAA hitters only had 19 qualified players in total, and we wanted to make all the datasets have the same sample size. We re-ranked these 19-observation datasets and took the top 5 players from each dataset. This allowed us to obtain 30 players (15 pitchers and 15 hitters) in total. These players and their data were then populated into two datasets: Final Hitters and Final Pitchers. The players in each of these datasets were then ranked by their total score.

Changes In Our Formula

To enhance the accuracy of our formula, we decided to move from a system that multiplied weights of each metric by the actual metric statistic, to multiplying the weights by a player’s relative ranking score in the sample. For example, for hitters, instead of multiplying our 40% weight for wRC+ by the players actual wRC+, we instead multiplied the 40% weight by how well they compared to other players in the sample. To explain this further, with a sample of 19 players from each divisional category (A+ pitchers, A+ hitters, AA pitchers, AA hitters, AAA pitchers, AAA hitters), the best person in each category for each metric would be awarded 19 points, and then have that multiplied by the weight of the metric, doing the same methodology for all 3 metrics for the hitters/pitchers. To illustrate, for AA pitchers, Matt Manning had the best xFIP out of the 19 pitchers, so he has awarded 19\*40% weight→ 7.6 points in his scoring. The reason we made this change was to “standardize” the statistics so there would be no bias in the data. When we used this methodology in week 1 for hitters for example, wRC+, which has an average metric of 100, was unintentionally more impactful in our scoring than the decimal statistic HR/FB. We believe by changing to a ranking system with an equal number of players in each sample (19), our final score will be unbiased and more representative of a player’s actual ability.

Another small change was making a pitchers final score be the higher the better exactly like hitters instead of the lower the better. This similar rank scoring methodology, with the same number of players in each sample, can allow pitchers and hitters to be compared to each other at the same time.

Passing Grade

Because our total score is calculated by multiplying weights to a player’s relative rank for each metric, our passing grade does not use a specific numerical score, but instead uses a percentile. We predict that whichever player has a total score (calculated by our formula) that is in the top 25% of the scores in a dataset (or equivalently, the 75th percentile) will be successful in their MLB career. This is a reasonable passing grade because you can see notable names from our final list of players. For example, in 2019 Matt Manning was the number 2 prospect in the Detroit Tigers franchise and number 27 prospect overall in the MLB. Another player from our final list is Dalton Varsho. Varsho was the number 5 prospect in the Arizona Diamondback system and number 100 prospect overall in the MLB in 2019. Another player is Dylan Carlson who was the St. Louis Cardinals number 1 prospect and the number 24 prospect overall in the MLB. Willi Castro, the 5th ranked hitter on our list, was an AL rookie of the year finalist last season as well.

Challenge Areas

One of the challenges we faced when identifying our thirty players was the fact that our formula is based on a rank system, which means that the weights that we assigned to the statistical metrics for hitters and pitchers were weighted based on the rank we assigned to them. For example, if a player received a rank of 14, then 0.4 (the weight assigned to the metric wRC+) would be multiplied by 14 rather than their actual wRC+ score. Since this is the manner in which our formula was created, it is important to note that our formula cannot be used until a person ranks a set of baseball players themselves and then utilizes our weighted formula since our formula would not work solely by multiplying weights to the metrics directly. Our formula would only work once a set of players has been ranked and the weights are assigned based on the rank itself.

Another challenge we faced was that the sample size for Triple-A (AAA) pitchers was small compared to the sample size of A+ and AA pitchers and hitters. The sample size for Triple-A (AAA) pitchers only contained 19 players after we filtered for age of players (between 18-25), which means that in order to make the scores consistent across the board, we had to condense our other sample sizes down to 19 as well. It is a known fact that the larger the sample size, the better; however, in this case, we did not have another choice. We wanted to make sure that the scores assigned to the players were consistent across the board in the sense that the final score consisted of the sum of weighted metrics multiplied by the same number of ranks. For example, hitters from A+, AA, and AAA with a rank of 19 would all receive a weighted score of 7.6 for wRC+. We wanted to make sure that the final scores for certain sample sizes would not be exponentially larger than the final scores for other sample sizes solely because certain sample sizes have more players since our formula is rank based and depends on the number of players within our sample size. Making the sample sizes all the same size made the data much more comparable and made it easier to identify 30 players.

Lastly, if we ran into a situation where certain players had the same final score, we would deal with it by ranking the higher division player higher in terms of ranking as well. For example, if two players both had a final score of 12.4 and one was in the AA division and the other was in the AAA division, the player in the AAA division would be ranked higher. However, if two players had the same final score, and were in the same division, then the player with the better individual metric rankings, with the effect of the different weights of different metrics taken into account, would be ranked higher.

Extra Notes

Please consider these factors when looking at our report and list of future players: The dataset we used was from the 2019 MiLB season since there was no 2020 MiLB season due to COVID-19. By using data from 2019, some players from our list could be out of the age restriction we set. In the beginning, we set a filter for only players from ages 18-25. With this dataset almost being 2 years old now, some players may be 27. Another thing to consider is that players in the minor leagues are always being traded, cut, or reassigned to a different league. Some players from our list can be on different teams or different leagues now in 2021 than stated on this list. The last thing to consider is that some of the players from our list may have made it to the MLB already. Due to the shortened MLB season last year, teams were consistently calling up players from their alternate site to fill in roster spots for other players when they had COVID-19 or were out due to contact tracing.

Follow List

Pitchers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Name** | **Age (in 2019)** | **Team** | **League** |
| 1 | Bryse Wilson | 22 | ATL | AAA |
| 2 | Kyle Wright | 24 | ATL | AAA |
| 3 | Matt Manning | 21 | DET | AA |
| 4 | Miguel Yajure | 21 | NYY | A+ |
| 5 | Paul Blackburn | 25 | OAK | AAA |
| 6 | Trevor Rogers | 21 | MIA | A+ |
| 7 | Kolby Allard | 22 | TEX | AAA |
| 8 | Mario Sanchez | 24 | WSN | AA |
| 9 | Thomas Jankins | 23 | MIL | AAA |
| 10 | Tommy Romero | 21 | TBR | A+ |
| 11 | Alec Bettinger | 23 | MIL | AA |
| 12 | Trey Supak | 23 | MIL | AA |
| 13 | Mac Sceroler | 24 | CIN | A+ |
| 14 | Daniel McGrath | 24 | BOS | AA |
| 15 | Maximo Castillo | 20 | TOR | A+ |

Hitters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Name** | **Age (in 2019)** | **Team** | **League** |
| 1 | Brian Mundell | 24 | COL | AAA |
| 2 | Daniel Johnson | 23 | CLE | AAA |
| 3 | Chris Gittens | 25 | NYY | AA |
| 4 | Luis Castro | 23 | COL | A+ |
| 5 | Willi Castro | 21 | DET | AAA |
| 6 | Tim Lopes | 24 | MIL | AAA |
| 7 | Abraham Toro | 22 | HOU | AA |
| 8 | Daulton Varsho | 22 | ARI | AA |
| 9 | Luis Campusano | 20 | SDP | A+ |
| 10 | Brennon Lund | 24 | LAA | AAA |
| 11 | Dylan Carlson | 20 | STL | AA |
| 12 | Lazaro Alonso | 24 | MIA | A+ |
| 13 | Luis Torrens | 23 | SDP | AA |
| 14 | Devin Mann | 22 | LAD | A+ |
| 15 | Casey Golden | 24 | COL | A+ |