John Incantalupo

MTH-370 Independent Study

May 6, 2024

<div align="center">How Cape League Performance Affects a Player's Draft Stock</div>

The MLB Draft is still an inexact science. Unlike with other major North American sports, the players drafted will not have an immediate impact on an MLB team's performance. A player will often spend multiple years in a minor league system before getting that fateful call to the show. This unique process drastically affects how scouts evaluate talent. They are looking for players with a teachable skillset who can become stars a few years down the road. With the introduction of sabermetrics as well as various ball tracking software such as the TrackMan camera system, scouts now have a whole new way of evaluating these draft prospects. Not only has TrackMan been used in all 30 MLB stadiums, but it has also made its way to some of the top collegiate leagues, such as the Cape Cod Baseball League.

Given that the Cape League is one of the top summer collegiate leagues in the country, it provides an adequate sample of many of the players who will hear their names called on draft night. Through our MTH-370 project, we sought to find a correlation between a Cape League player's wins above replacement and his draft position. We also aimed to see if there were any TrackMan metrics that were highly correlated with draft position in order to gain a sense of what made particular players so highly touted.

At our disposal, we have the play-by-play and TrackMan data from both the 2022 and 2023 Cape League seasons. We also have the results from the 2023 MLB draft, which features many of the stars from the 2022 Cape League season. The 2022 Cape League data combined

with the 2023 draft results will serve as our training set, so that we can use the 2023 Cape

League data to predict where players will go in the upcoming 2024 MLB draft.

Once we had compiled each drafted player's WAR total and compared it to their draft

position, we found that position players have a better correlation between the two variables than

pitchers. In fact, the pitcher with the most Cape League WAR in 2022, Bryce Warrecker, was not

taken until the 20th round and was the last Cape League pitcher taken. One possible reason for

the lower correlation for pitchers is the recent attention given to their workload, which would

make sense since WAR is a cumulative statistic that builds up over time. Although we have

shown that WAR alone is not the best predictor of MLB draft position, we now have access to

TrackMan data from the past two Cape League seasons. We can test to see if there are any

metrics from TrackMan that correlate highly with draft position. Unfortunately, the TrackMan

data that we received required a lot of cleaning and reformatting. The main issue that arose was

the fact that the player names from the TrackMan data were stored in a different format than the

PointStreak play-by-play data. This was problematic since we wanted to merge the two datasets.

Thanks to Tableau Prep, I was able to clean up all the typos and misspellings in the TrackMan

data. Then, I split both of the singular player name columns into two columns for the first and

last name. By using an R program, the player names from the PointStreak data were also split

into two columns, and the two datasets were then merged by pitch.

From the TrackMan data, a few of the main metrics that we looked to include in the

model were barrel rate and exit velocity for position players, and whiff rate and fastball spin rate

for pitchers. We also wanted to include framing data for catchers but scrapped it due to the

limited number of catchers in the dataset. After testing various types of models, such as a linear

model, random forest, elastic net, and ridge regression, we found that the linear model

surprisingly gave us the lowest RMSE and MAE. When analyzing feature importance, WAR was second in both the position players' and pitchers' models to exit velocity and whiff rate, respectively. This makes sense, as hitting the ball hard as a hitter and generating swings and misses as a pitcher are two important skills that many scouts today are looking for.

Although the linear model produced the lowest RMSE and MAE on our training data, it was not necessarily the best at predicting the 2024 MLB draft. For the hitters, the lowest predicted draft pick was Travis Bazzana, who the model says will go at pick 90, which is the middle of the third round. Many online mock drafts have Bazzana going as high as the first overall pick, so there is a pretty large discrepancy between our model and other draft projections. The predictions for the pitchers were even worse, as some of the predicted values were less than 0, which is impossible. One possible reason is overfitting. Simply having a low RMSE and MAE does not imply that the model will be a good fit on other datasets, and this rings true with our linear model. However, this does not mean that our linear model is useless. Although Travis Bazzana will definitely be drafted much higher than pick 90, the model's claim that he will be the top position player selected is a pretty reasonable prediction. In the future, this model will be better suited as a predictor for the order of players to be drafted rather than actually predicting the pick number.

At this point, all we have to do is wait. The 2024 MLB draft is less than three months away. Dozens of players who took part in Cape League action last summer will have their lives changed forever with one simple phone call. As for us, we will be looking to see how accurate our linear model was in predicting which former Cape League players would be drafted first. Just as we saw with the influx of TrackMan data over the past several years, the way that scouts will assess future talent will look much different in a decade. Perhaps more advanced metrics such as

bat speed and spin axis will be the new buzz words in baseball circles. Regardless, it is certainly going to be interesting to see what comes next.