Will He Sign or Will He Go (to college)?

An analysis of high school players in the Major League Baseball Entry Draft

Eli Gnesin

May 1, 2020

1 Introduction

Held every year in June, the Major League Baseball (MLB) First-Year Player Draft functions as the entry point for many young players into professional baseball. Typically spanning 40 rounds, teams take turns, in reverse order of their standings from the previous season, to select players [1]. Teams also can receive picks by extending qualifying offers to eligible players and then having those players sign elsewhere, or through "Competitive Balance", which helps benefit clubs with lower revenue and clubs in smaller markets [2] [3]. Likewise, teams can lose draft picks as punishment for various infractions or by signing players who rejected qualifying offers to play elsewhere [2].

During its pick, a team can draft any player from the pool of eligible players. This pool includes any player who is resident of the United States or Canada or any US Territory, such as Puerto Rico [1]. However, there are only three groups of players who are eligible [1]. These are:

1. High School players who have graduated but not attended college
2. Junior College players under all circumstances
3. 4-Year College players who are 21 years old or have finished at least their junior year

Once a player is drafted, the player and team negotiate a contract for the player to join the organization. Players with remaining eligibility can, however, choose not to sign, and instead return to, or enter, a 4-year or junior college. In these cases, the player becomes eligible to be drafted again in future drafts, so long as they fall within the eligibility criteria [1].

The goal of this analysis is to examine the subset of High School players drafted in the MLB draft. The first objective is to examine the various characteristics of this data for trends in whether or not these players have signed. The second objective is to build a model that can predict whether or not a high school player drafted with certain characteristics will sign his contract. The final objective is to apply this model to the 2020 MLB draft and examine the results from it.

2 Data

The data set for this analysis was collected from the Baseball Reference draft search engine at www.baseball-reference.com. I collected data for the first 20 rounds, and all supplemental rounds, of every draft between the years 2007 and 2019. The start date of 2007 was chosen because it was the first year of a Collective Bargaining Agreement (CBA) [4]. Once concatenated into a single data frame using python, the dataset spanned 8080 rows and 23 columns, with a subset of 2137 high school players. I cleaned the data of anomalies in the position column and combined the "supplemental" rounds with their non-supplemental counterparts. Certain rows in the data required manual updating of missing data in the columns related to Position, Signing status, and "Drafted From", which indicates the school where a player was drafted out of. I also wrote short functions to create a column for the state and region from which a player was drafted and the CBA during which he was drafted.

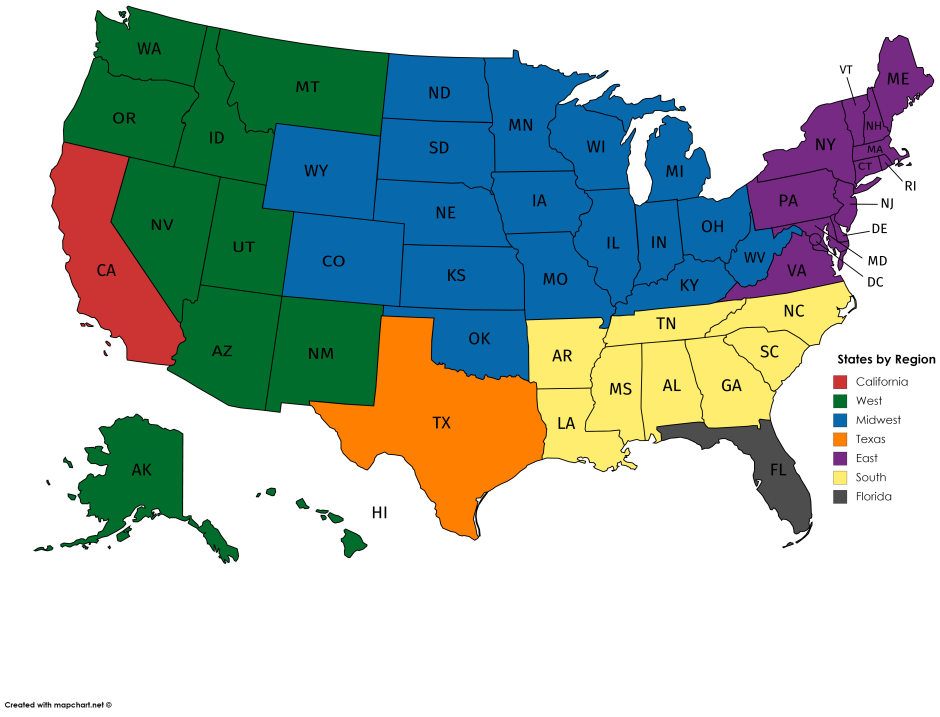

Figure 1: The Map of Regions for the Regional Split

Figure 1 shows most of the regions used for this analysis. Regions were chosen to avoid overly unbalanced classes while preserving some regional agreement. Texas, Florida, and California were separated because they each had enough players to create their own regions. The only region not included in Figure 1 was a region for players drafted outside the 50 United States, which included players drafted from Canada, Puerto Rico, and other United States Territories, as well as a small group of players drafted from elsewhere.

3 Part 1: Exploration

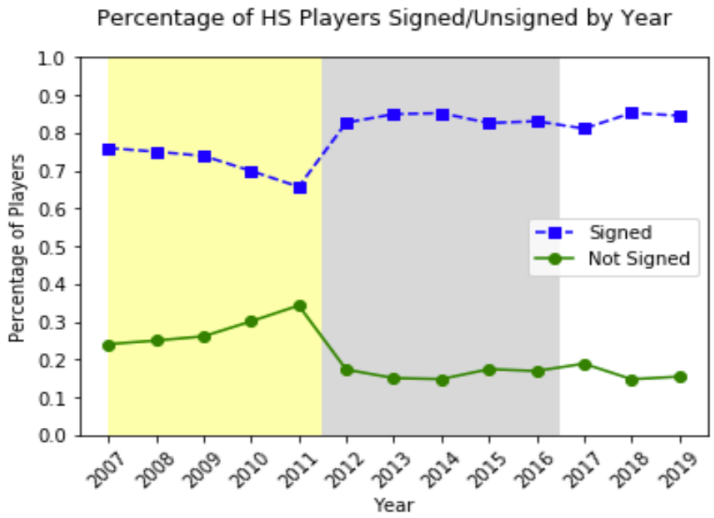

Figure 2: The graph of percentage of players who signed by year

Figure 2 shows the percentage of drafted high school players who signed and did not sign by year. A lower percentage of high school players signed during every year between 2007 and 2011 than in any year after. This period (2007-2011), correlates to the first Collective Bargaining Agreement in my dataset, and is shaded in yellow. After the 2011 season, the new Collective Bargaining Agreement changed the way contracts were handled with regards to the amount of money teams could spend to sign players in the first 10 rounds and in subsequent rounds [4]. This second Collective Bargaining Agreement lasted through 2016 and is shaded in gray on Figure 2.

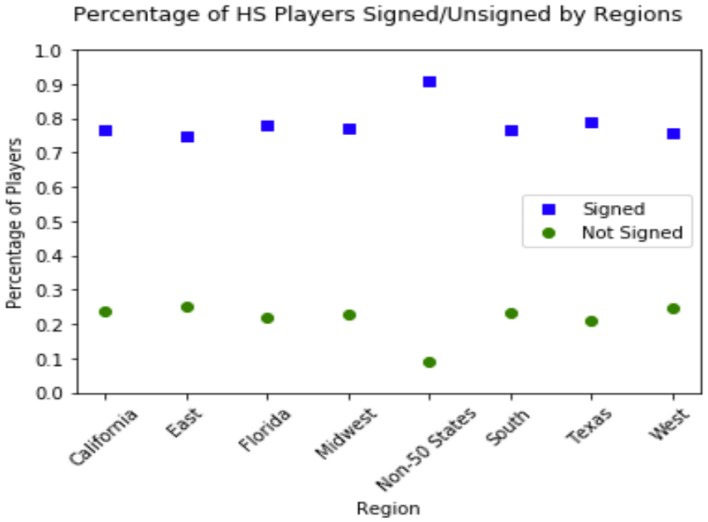

Figure 3: The graph of percentage of players who signed by region

Figure 3 shows the percentage of drafted high school players who signed by region. On the whole, there was very little variation in the percentage of players who signed by region, with the notable exception being players drafted from outside the 50 states. This group, primarily included players from Canada and Puerto Rico, but it also had a couple of players each from Cuba, Germany, and the United States Virgin Islands, chose to sign at a significantly higher rate than any other group, with about 91% of players signing. This stems from over 70% of the region's players coming from Puerto Rico, and approximately 95% of those Puerto Rican High School players signing.

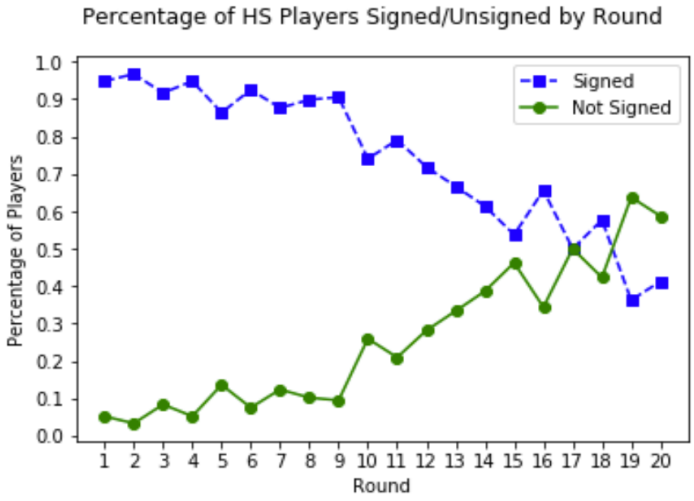

Figure 4: The graph of percentage of players who signed by round

Figure 4 shows the percentage of drafted high school players who signed by the round they were drafted in, with supplemental rounds grouped into their main rounds due to lack of sample size. There is a slight downward trend in the percentage of players who sign in the first 9 rounds, followed by a much steeper downward trend in later rounds. This culminates in the majority of high school players drafted in rounds 19 and 20 choosing not to sign. Importantly, there are 40 rounds currently in the MLB draft, and there were 50 rounds between 2007 and 2011 [4]. Likely, these trends would continue, such that in the later rounds of the draft, the vast majority of high school players chose not to sign.

4 Part 2: Modeling

After graphically analyzing trends and anomalies in these characteristics, I moved to train a model that could accurately say whether a high school player with a given set of draft characteristics chose to sign his contract or not. I first took a subset of the high school player data. This subset had columns for their draft round, the collective bargaining agreement in effect when they were drafted, their position, region, the team that drafted them, and their signing status. I then realized that, given 30 MLB teams and a minimum of 10 positions (the 8 field positions and both left-handed and right-handed pitchers), there would be far too many binary quantitative variables to render a model that was not over-fit.

As such, I went back to the initial data and created two more columns. The first column was a "Position Group" column, which separated players into four groups: infielders, outfielders, pitchers, and catchers and utility players. The second column was a "Division" column, which separated players based on their teams into the 6 divisions in which those teams play [5]. This reduced the 40 binary quantitative variables for position and team into 10 for position group and division. From there, I used a column transformer to encode the variables for division, position group, region, and CBA into binary quantitative variables, which gave 22 independent variables.

The first model used was a simple Logistic Regression model. I used train and test subsets with a test set size of 20% and a random state of 1693 to allow for ease of replication. A logistic regression model fit to this training set with the "lbfgs" solver returned a precision of 85.8% and recall of 93.7% on the test set with size of approximately 428 players. I then used a Support Vector classifier, a Naive Bayes classifier, and a Nearest Neighbors classifier on the same test and training set, but with lower precision and recall on the test set.

The final set of classifiers I tried was TPOT, or the Tree-based Pipeline Optimization Tool. TPOT is an automation algorithm that iterates through thousands of possible pipelines with a variety of different classification models and data processing techniques, all with cross-validation, and then returns the strongest pipeline in terms of one of many scoring techniques [6]. I ran multiple instances of TPOT with different population sizes, different numbers of iterations, and different learning rates, but no instance returned a better model than the original logistic regression model.

5 The 2020 Draft and COVID-19

Once the logistic regression model was chosen, I moved on to my final objective. When this analysis was conceived, the plan was to craft a model based on the 2007-2019 MLB drafts and then apply that model to the 2020 draft in real time and present the results. However, in May 2020, due to the Covid-19 pandemic, Major League Baseball chose to shorten the draft from 40 rounds to 5 rounds [7]. After the draft took place, I collected data from all five rounds using the MLB draft search engine on Baseball-Reference. In total, this data set contained 160 players and included a subset of 47 high school players.

After collecting the data set, I processed it using the same pre-processing, adding the same columns as with the whole dataset. I then separated out the signed values and encoded the discrete independent variables into binary quantitative variables, as before. Using the whole subset of 47 high school players as a test set, the logistic regression model was applied to the data to generate prediction values, which were then compared against the actual signing values for the 47 players to determine how strong the model was. The logistic regression model, which returned nearly 86% precision on test sets, returned only 70% precision on this 2020 test set. To look deeper, I printed the predicted values and the target values for each of the 47 players:

```
The actual values for the 2020 draft class:

[1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 0. 1. 0. 1. 1. 1. 1. 1. 0. 1. 1.
 1. 0. 0. 1. 0. 1. 0. 1. 0. 0. 1. 0. 1. 1. 0. 0. 1. 1. 0. 1. 1. 1. 1.]

The logistic model predicted values for the 2020 draft class:

[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
```

Figure 5: The printouts for the logistic model on the 2020 MLB draft data

Figure 5 shows the printout for the actual signing values and logistic model predicted signing values for each of the 47 high school players drafted in the 2020 MLB draft. A 1 in either of the arrays indicates a player who signed, and a 0 indicates a player who did not sign. The string of only 1's for the logistic model predicted values indicates that the model thought every high school player drafted in 2020 would sign his contract. To determine if this was a fluke, I repeated this process with each of the other models created in Part 2, with the same results in every case.

I then considered whether the issue was that the model had not been fit on only the first five rounds, and whether that would change anything. Thus, I repeated the process in Part 2, but with my dataset used for training containing only the first five rounds of every draft from 2007-2019. This, however, proved to be fruitless, as all the models fit on the first five rounds still predicted that every high school player in the 2020 MLB draft would sign his contract.

6 Discussion

Throughout this analysis, there were several trends and oddities that stood out and that I believe would be worth further exploration. The first of these observations was the high percentage of Puerto Rican High School players who sign their contracts, which was almost 95%. Further research could explore the potential cultural or socioeconomic reasons behind this. Another observation that could be explored was the sharp decrease in the percentage of players who sign after round 9 of the MLB draft, as seen in Figure 4, and the potential reasons for such a decline in the context of how the MLB draft money is allocated or changes in the draft over time. A final avenue which future research could pursue is the effect of other factors, beyond the shortening to 5 rounds, on the 2020 draft. This could include financial or social impacts of the pandemic, as well as financial impacts on Major League Baseball and Minor League Baseball. The existence of other factors could explain the discrepancy between the model and the actual resul

7 Conclusion

I have graphically analyzed some of the characteristics related to players drafted in the MLB draft, and I have fit a model to predict whether or not a high school player who is

drafted will sign his contract. I have also attempted to apply this model to the 2020 MLB draft, though with poor results. Further research could be done to expand the data set to earlier drafts and to expand further to include 30 or even 40 rounds from each draft. This analysis could also be expanded to include college and/or junior college players.

References

[1] MLB first-year player draft rules. http://mlb.mlb.com/mlb/draftday/rules.jsp.

[2] Anthony Castrovince. The qualifying offer rules, explained, October 2017. https://www.mlb.com/news/mlb-qualifying-offer-rules-explained-c259650658.

[3] Competitive balance draft picks. http://m.mlb.com/glossary/transactions/competitive-balance-draft-picks.

[4] CBA history. Baseball Prospectus, https://legacy.baseballprospectus.com/compensation/cots/league-info/cba-history/.

[5] MLB divisions. https://www.mlb.com/standings.

[6] TPOT. https://epistasislab.github.io/tpot/.

[7] Jeff Passan and Kiley McDaniel. Sources: Mlb shortens 2020 draft from 40 rounds to 5, May 2020. https://www.espn.com/mlb/story/_/id/29152007/mlb-shortens-2020-draft-40-rounds-5.