Analyzing MLB Draft Biases and Modeling Player Success using Wins Above Replacement

STATS 141XP Final Project

Peter DePaul, Anish Ravilla, Robin Lee
Kevin Kim, Alan Wong, Hongye Zheng

2024-03-15

Abstract

In Major League Baseball, the First Year Player Draft is the primary system for teams to select amateur baseball players from high school, colleges, and other amateur baseball clubs. With this report, we seek to not only identify attributes and physical characteristics that influence a draftee's selection, but also develop a model to predict who ultimately succeeds. We explore recurring trends among drafted players—such as their patterns in home state, prior experience, and time in the MLB—and analyze their relationship with Wins After Replacement. By quantifying success as a player's overall WAR, we also construct a Gradient Boosted Decision Tree model for predicting WAR. Our model tends to accurately predict which players perform poorly in their careers, but struggles to predict players with average and high performing careers. Furthermore, we discover that there is some bias favoring players from college and certain home states which may affect their draft potential. While there is some statistical evidence suggesting a player's position and debut age correlates with their WAR, we have found little success predicting a player's WAR using their draft information and demographics. Ultimately, the charm and excitement of Major League Baseball is in its unpredictability—where success and talent is not defined by a quantifiable attribute or a specific set of traits.

Contents

1 Introduction

In the world of professional baseball—specifically Major League Baseball (MLB)—the First Year Player Draft is a significant moment where teams meticulously evaluate and select amateur baseball players from high schools, colleges, and other amateur baseball clubs. However, even though the selection process should technically be based on objective and impartial reasoning, we are confident that various multifaceted biases subtly influence the decisions made by scouts and team executives (Caporale and Collier 2013).

1.1 Background

The baseball draft is a process in which the MLB's 30 professional baseball franchises gather together to draft amateur players, which mainly consist of players from the high school and college circuits (Staudohar, Lowenthal, and Lima 2006). Teams take turns picking players for their roster. Keep in mind that some players are drafted more than once. In this case, players are usually drafted out of high school, but they decide not to sign and instead play college baseball; they re-enter the draft after completing college.

Teams are awarded draft picks based on a **draft lottery**, since 2023, with the teams that did not make the playoffs the previous year being entered into the lottery, and are given the opportunity to go first in the draft. The draft currently undergoes 20 rounds of selections, where each team gets to pick a player of interest for their roster. This draft, occurring in June, is also known as the **First-Team Player Draft**, where players who enter are typically high school or community college/four-year college graduates, and have never played for any professional baseball team (Garmon 2012).

1.2 Variable Overview

By implementing statistical data analysis, we seek to investigate how a scout's perception of talent and potential is influenced by a prospective player's attributes—such as their age, position, dominant hand, and educational background (Conforti, Crotin, and Oseguera 2022). Furthermore, in order to quantify the success of drafted players, we utilize their **Wins Above Replacement (WAR)**. For a comprehensive overview of all parameters, see Table 1. We accessed our data from Bill Petti's BaseballR package.

We hope to identify biases that influence when a player is drafted, and determine whether these decisions are meritorious based on said player's performance (Crotin et al. 2023).

Label	Description	Unit of Measure
fg_playerID	Player ID	Numeric
Name	Player name	Character
fWAR	Wins Above Replacement	Numeric
pick_round	Which round a player is picked	Numeric
pick_number	Which number a player is picked	Numeric
year	Year a player is picked	Numeric
person_birth_state_province	State or province the player is born	Character
person_height	Height of player	Numeric
person_weight	Weight of player	Numeric
person_primary_position_abbreviation	Player's primary position	Character
person_bat_side_code	Player's batting side (L/R)	Binary
person_pitch_hand_code	Player's pitching hand (L/R)	Binary
mlb_played_first	Year of first MLB game	Numeric
mlb_played_last	Year of last MLB game	Numeric
high_school	Player went to high school	Binary
home_state	Player's home state	Character

Table 1: Variable Overview

2 Exploratory Data Analysis

2.1 Relationship between drafting High School or College Players

In order to analyze the relationship between draft round and high school status, we created a barplot showing the percent of draftees by round with colors denoting high school status when drafted. By analyzing this barplot [see Figure 1], we can see that as you get further into the first 10 rounds that siginficantly less high school players are drafted.

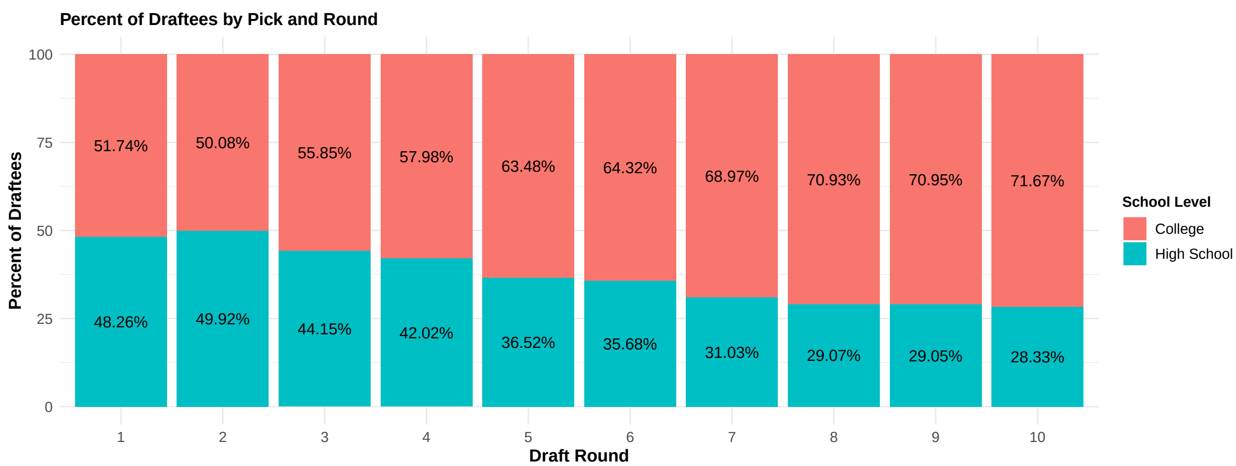

Figure 1: Percentage of Players by Round and School Level

2.2 Histogram of Time to Play First MLB Game

In order to analyze the number of years before a draftee played their first MLB game, we calculated the number of years between a player's first MLB game and the year they were drafted.

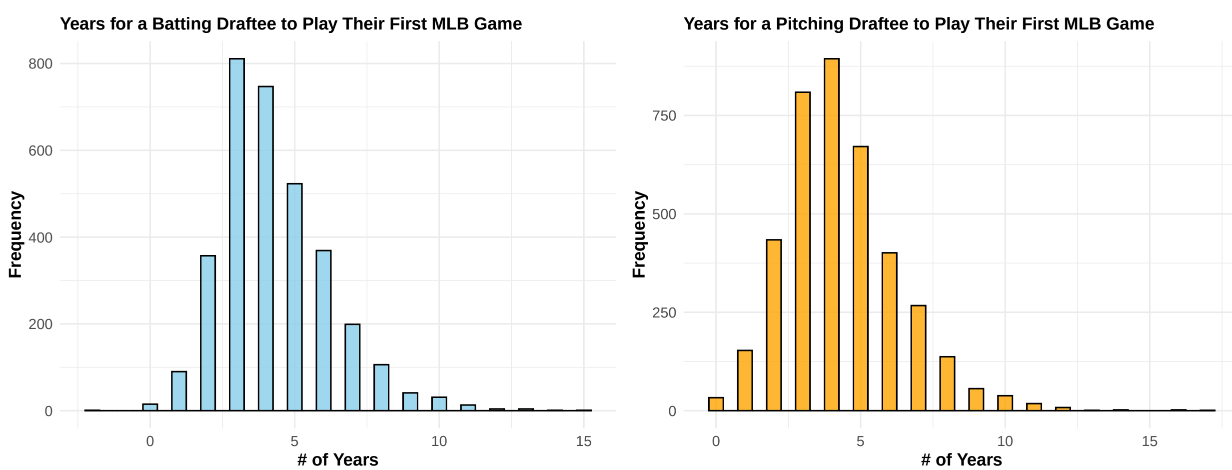

Figure 2: Histograms of Years before First MLB Game

By analyzing these two histograms [see Figure 2], the number of years for a batting draftee to play their first MLB game is similar to the number of years for a pitching draftee. Both datasets are slightly right-skewed, with most people playing their first game between 3-5 years after their draft.

2.3 Relationship between Weight and Height

We also investigated the relationship between weight and height for both batting and pitching draftees. By creating a scatterplot of both datasets, we not only mapped each individual person's attributes, but also plotted the mean weight/height and performed a simple linear regression to determine the overall trend.

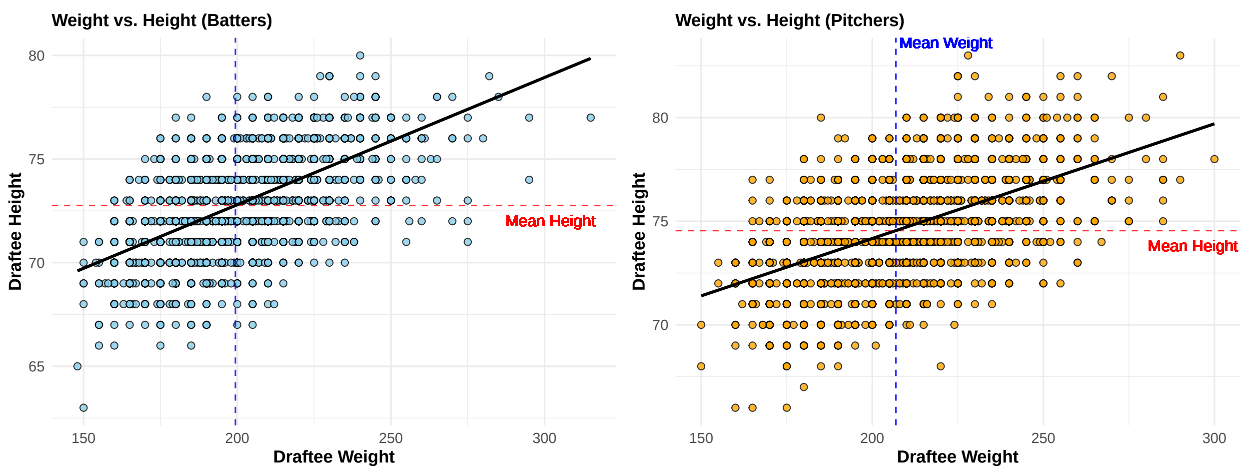

Figure 3: Scatterplots of Mean vs. Weight

Comparing these two plots [see Figure 3], the mean weight for batting draftees is slightly smaller than the mean weight for pitching draftees. Likewise, the mean height of batting draftees is also slightly smaller than the mean height of their pitching counterparts. We can expect the average batting draftee to be slightly shorter and lighter than the average pitching draftee.

2.4 Correlation Heatmap

The correlation between pick round, year, height, weight, and WAR for both player categories is:

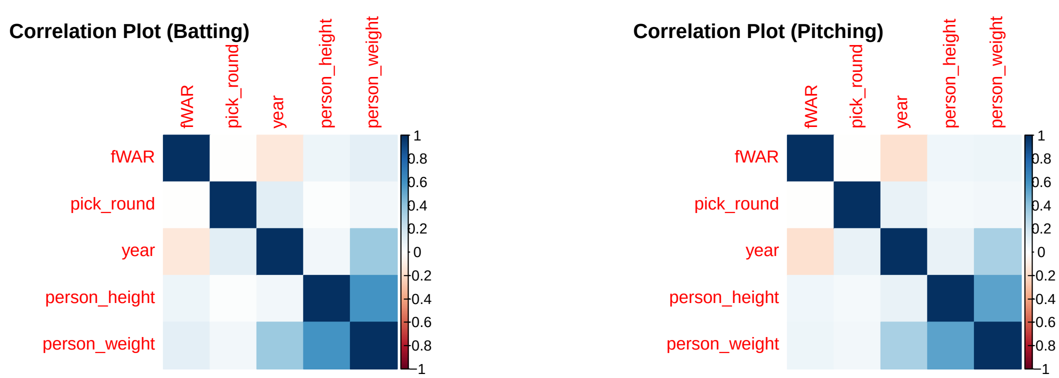

Figure 4: Correlation Plots for Pitching and Batting

As expected [see Figure 4], there is a strong correlation between a person's height and their weight. For pitching, there is a small positive correlation between a player's height and their WAR; for batting, there is a negative correlation between a player's height and their WAR. We see there is no correlation between a batting draftees' weight and WAR, but there is a slight positive correlation for a pitching draftee.

2.5 Home State Frequency Heatmap

In order to investigate which state produces the most draftees, we construct a frequency heatmap:

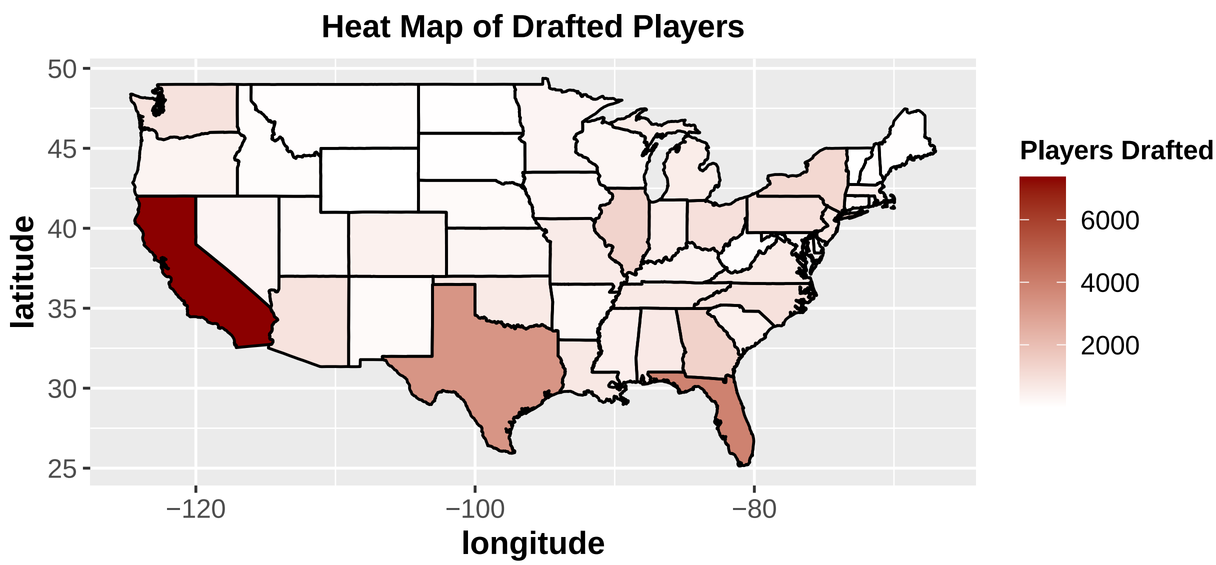

Figure 5: Frequency Heatmap of Players per State

From our state map [see Figure 5], we can see that most players drafted are originally from California, with Texas and Florida as the other two primary home states. We can attribute this trend to weather, as these states tend to have higher temperatures and little-to-no snow compared to East, Midwest, and Northwest, which allows people to play baseball all year long and for youth leagues to have more flexibility with their seasons. These states also have relatively large populations, which could easily explain why the most draftees emerge from these states.

2.6 Relationship between Time in MLB and WAR

Exploring how the length of player's career affects their WAR, we construct the following scatterplot:

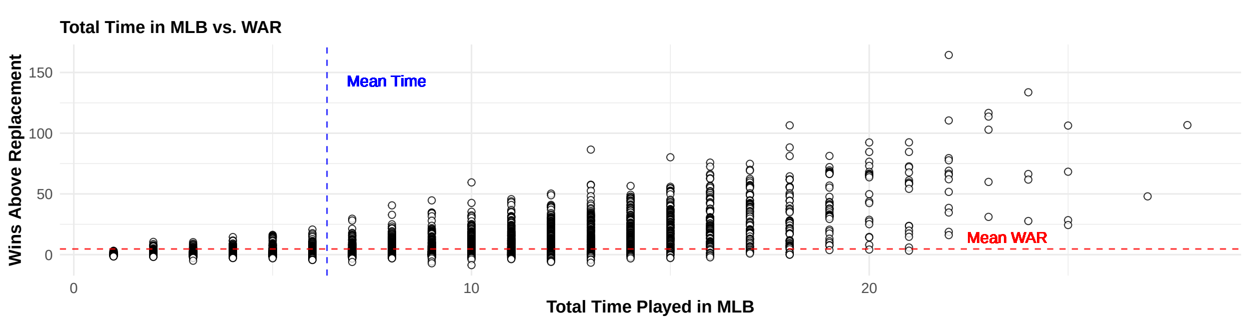

Figure 6: Relationship between Time Played in MLB vs. WAR

By comparing each player's total time played in MLB and WAR [see Figure 6], we can see that the length of a player's career does not always equate to a high WAR. However, there does seem to be an upward trend, which makes sense as we expect a player to win more as they play longer.

3 Analysis and Model of Player Success

3.1 Who is Successful?

3.1.1 How often do Draftees make it to the MLB?

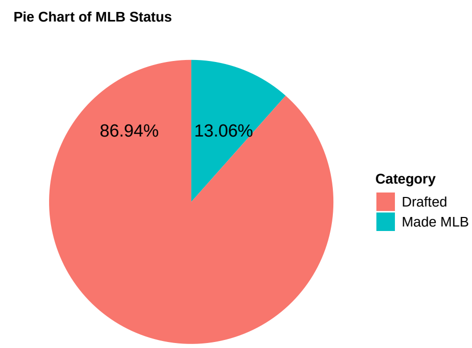

Figure 7: Pie Chart of Draftees MLB Status

By looking at the pie chart of whether or not draftees ultimately made it to the MLB [see Figure 7], we can see that only 13.06% of players actually received the opportunity to play and a majority of players miss the major league. We will focus specifically on the players who made it to the MLB.

3.1.2 What is the Best Draft Class in MLB History?

We can see from the barplot [see Figure 8] that the 1965, 1985, and 2002 draft classes are highly successful with each garnering over 900 total fWAR across the draft class. 1965 includes the likes of Johnny Bench, Nolan Ryan, and Tom Seaver. Meanwhile 1985 includes Barry Bonds, John Smoltz, and Randy Johnson. 2002 includes Zack Greinke, Prince Fielder, and Cole Hamels. The thing these drafts share in common is that they have an abundance of talent, both in terms of Hall of Famers and depth.

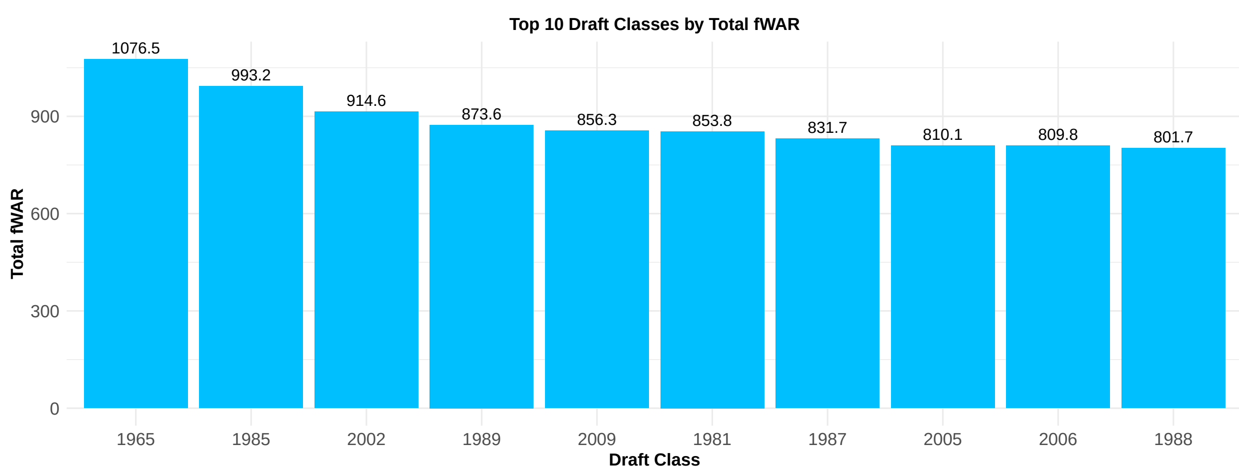

Figure 8: Best Draft Classes by fWAR

3.1.3 What Positions have been Successful?

We are able to see that the positions which are most successful have often been outfielders (CF or LF), First Basemen (1B), or Third Basemen (3B) [see Figure 9]. This can be explained by First Basemen possessing game changing amount of power hitting, and the high level fielding roles of outfield and third base.

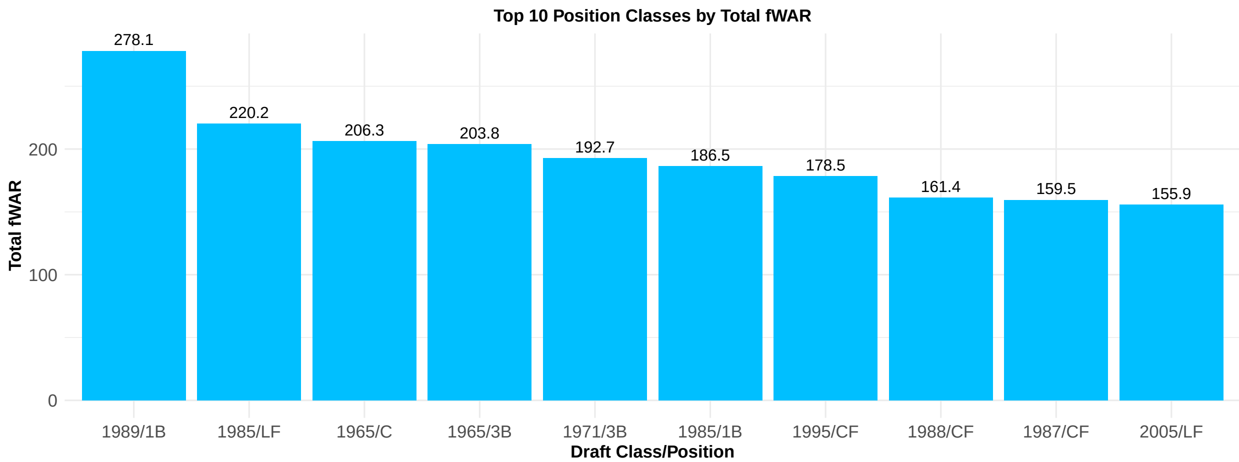

Figure 9: Best Draft Classes by Year and Position

3.2 Prediction Model

By defining each player's fWAR as the definition of success, we aim to create a classification model which predicts a players' categorical level of success only from their physical characteristics, their demographics, and the characteristics of their draft position.

3.2.1 Feature Engineering

We believe it would be difficult to predict a player's specific fWAR numerical value without any informational statistics of their playing careers. Thus, in order to make this a simpler process, we have come up with the idea of creating an ordinal categorical (factor) variable named `fWAR_class` with 6 levels:

- Greater than 25 fWAR
- 20-25 fWAR
- 15-20 fWAR
- 10-15 fWAR
- 5-10 fWAR
- Less than 5 fWAR

Additionally, in order to assess the impact of where in the draft a player is picked and its affect on their WAR, we add several binary variables indicating if the player was any of the following:

- Player was the #1 overall pick in the draft (`top_pick`)
- Player was a top 10 overall pick in the draft (`top_10`)
- Player was a top 25 overall pick in the draft (`top_25`)
- Player was a top 50 overall pick in the draft (`top_50`)
- Player was a top 100 overall pick in the draft (`top_100`)

We have also added a few more descriptive characteristics of the draftees including:

- `yrs_before_debut` which is the number of years before a player debuts in the MLB
- `draft_age` which is a player's age when they are drafted into the MLB
- `debut_age` which is a player's age when they made their MLB debut

3.2.2 Feature Selection

Following our feature engineering process, we employ the Boruta feature selection algorithm in order to ensure we are not using unnecessary variables in our modeling process. This step will strengthen the predictive power of our model. The Boruta algorithm is used to confirm the importance of variables by implementing Random Forest methods based upon a given classification formula.

The Boruta process confirms that all predictive features in the data set are important.

3.2.3 Model Creation

For our model, we decide to utilize the XGBoost library in order to create a gradient boosted decision tree for predicting `fWAR_class`. A boosted tree is built in sequential order such that it is able to learn from previous iterations for classification purposes. On top of this, boosted tree can be tuned to prevent overfitting through the use of tree depth and penalty functions. We utilize the tidymodels library as well in order to simplify our workflow and recipe process. Our step-by-step procedure is as follows:

1. We split our overall data into a training (70%) and testing (30%) data sets stratified by player position.

2. To take it a step further, we utilized the hyperparameter tuning process for our boosted decision tree in order to find the ideal parameters—such as learning rate and tree depth—for our data.

3. As part of our model process (recipe and pre-processing), we check to ensure no variables have zero variance, add dummy variables for all categories which are not already encoded as such, and make sure to normalize all numeric predictors such that they are on a similar scale. These choices minimize the variance caused by outliers and differently categorized variables.

4. We perform our model tuning process by generating a random grid of parameters and continually testing each hyperparameter under a different set of values. By measuring their performance, we continually update the hyper-parameters to the ones with the highest classified "accuracy" measure.

Ultimately, our Boosted Decision Tree consists of the following 8 tuning parameter values [see Table 2]:

Parameter Description	R Documentation	Value
Number of Randomly Selected Predictors	mtry	17
Number of trees	trees	98
Minimal Node Size	min_n	8
Tree Depth	tree_depth	9
Learning Rate	learn_rate	≈0.09885
Minimum Loss Reduction	loss_reduction	≈0
Sample Size	sample_size	1
Number of Iterations Before Stopping	stop_iter	5

Table 2: Boosted Tree Tuning Parameters

From there, we perform a 10-fold cross-validation on our Boosted Tree model to assess the stability of our parameters and ensure the performance is consistent. As such, we randomly partition the original sample into k equal-sized subsamples (folds), before testing and training the model on each different iteration.

4 Results

4.1 Cross-Validation Metrics

After performing 10-fold cross-validation, we are able to measure the accuracy of our model and determine its predictive power by constructing the following metrics table. This does not specify the model's ability to predict each individual WAR category, but rather computes the overall accuracy for all classes.

Table 3: Table of Cross-Validation Metrics

metric	estimator	mean	folds	standard_error
accuracy	multiclass	0.7832	10	0.0056
f_meas	macro	0.5345	10	0.0143
roc_auc	hand_till	0.6076	10	0.0055

According to our cross-validation metrics [see Table 3], our model perform moderately well. This is because an accuracy of 78.32% is considered a "moderately accurate" model, though we believe this is mostly caused by our model's difficulties in distinguishing between classes. The `roc_auc` score further shows that the model struggles to distinguish between the positive and negative classes, as it is almost a 50-50 chance.

4.2 Prediction Metrics by Class

Focusing on our models ability to predict each individual WAR category, we calculate our model's sensitivity and detection rate for the respective classes, alongside several other metrics.

Table 4: Prediction Metrics by Class

	Sensitivity	Specificity	F1	Prevalence	Detection Rate	Detection Prevalence
Class: Greater than 25 fWAR	0.155	0.984	0.217	0.055	0.009	0.024
Class: 20-25 fWAR	0.000	1.000	NA	0.020	0.000	0.000
Class: 15-20 fWAR	0.000	1.000	NA	0.025	0.000	0.000
Class: 10-15 fWAR	0.000	1.000	NA	0.042	0.000	0.000
Class: 5-10 fWAR	0.000	1.000	NA	0.083	0.000	0.000
Class: Less than 5 fWAR	0.991	0.074	0.877	0.774	0.767	0.976

Analyzing our model's prediction metrics according to each WAR category [see Table 4], our model performs best when predicting players with less than 5 fWAR and greatly suffers as it increases. In other words, it is especially good at identifying players who will perform below average (as seen by 0.991 Sensitivity and 0.767 Detection Rate). However, our model suffers when predicting players who are average or better, as their detection rates are all 0%, with no detection prevalence and sensitivity. There is some success when predicting players with greater than 25 fWAR, but because the detection rate is less than 1%, we do not believe that our model is very accurate when identifying this category either.

4.3 Confusion Matrix of Model Predictions

Because we are utilizing a classification algorithm, we construct a confusion matrix to summarize the performance of our model, specifically when predicting a player's success (WAR). The confusion matrix shows the number of correct predictions, as expressed by the diagonal values. Non-diagonal values are the number of incorrect classifications. The rows signify the actual values, while the columns indicate the predicted values. We use this table to identify which WAR categories our model classifies correctly or incorrectly the most.

Table 5: Confusion Matrix of Predictions

	Greater than 25 fWAR	20-25 fWAR	15-20 fWAR	10-15 fWAR	5-10 fWAR	Less than 5 fWAR
Greater than 25 fWAR	18	2	4	6	5	15
20-25 fWAR	0	0	0	0	0	0
15-20 fWAR	0	0	0	0	0	0
10-15 fWAR	0	0	0	0	0	0
5-10 fWAR	0	0	0	0	0	0
Less than 5 fWAR	98	39	49	83	169	1608

From our confusion matrix of model predictions [see Table 5], our model correctly classified 13 players as greater than 25 fWAR and 1605 as less than 5 fWAR. However, it incorrectly classified 36 players who have a fWAR greater than 25, and incorrectly classified 442 players who have a fWAR less than 5. Overall, our model works best for predicting players with less than 5 fWAR, and signficantly lacks when predicting players with 5-25 fWAR. There is some success for players with more than 25 fWAR, but it is not accurate.

4.4 Variable Importance Plot

Finally, we seek to identify which predictors are most important when predicting a player's success or WAR. Using the results from our Boosted Tree model, we rank each feature in the prediction process, and identify whether or not they have a strong impact on a player's WAR.

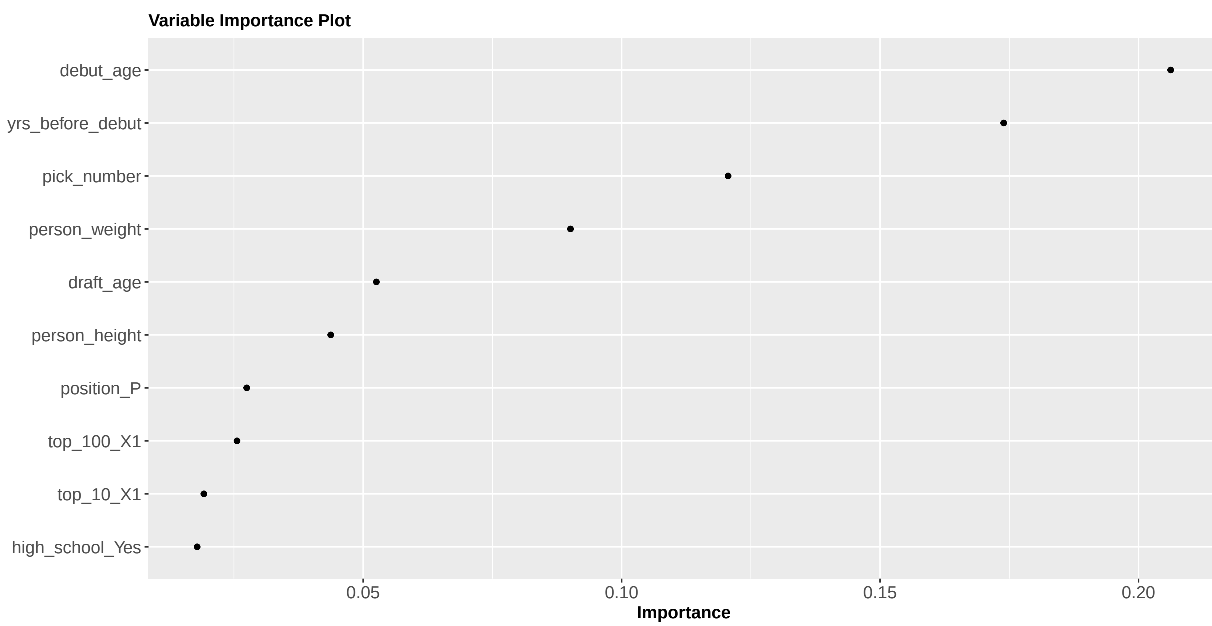

Figure 10: Variable Importance Plot

According to the variable importance plot [see Figure 10], a player's debut age has the greatest importance; this is followed by the number of years before their debut, their pick number, their weight, height, age when drafted, and position.

These attributes are generally expected, as the age of a player at debut and when they were drafted is often indicative of how much practice they had before entering the major leagues. Furthermore, we generally expect players who are drafted earlier to perform better, and certain physical characteristics certainly affect their performance. Finally, we believe that the position of a player influences their success [see Section 3.1.3 for further information].

5 Conclusion

In this report, we are able to discover several important characteristics when identifying MLB draft biases:

5.1 High School vs. College Baseball Players Bias

The first of these biases is the bias towards drafting high school vs. college baseball players. More often than not college players are drafted more frequently, but this is likely due to the fact that teams are more cautious with drafting high school players unless they're exemplary. This is illustrated by the higher rate of high school players being drafted in the rounds 1 through 5 than anywhere else in the draft. This supports the idea that the high school players being drafted are elite, and are being drafted very early on in the draft.

5.2 Home State Bias

Another factor we were able to unlock was the draft bias for players depending upon their home state. To no surprise players were most frequently from larger states such as California, Texas, and Florida but all states have high populations. Aside from this there was an affinity for players hailing from states in warmer climates such as states in the south, and states in the west. It's likely teams would be hesitant to draft players from an underrepresented state like Vermont, or Wyoming due to the lower talent pool, and the lack of history and certainty of players from those areas historically.

5.3 Relationship between Player Position and WAR

Using our metric for success as `fWAR` we were able to determine from our statistical data concerning draft prospects who made the MLB that those players who tend to succeed most in their careers are those playing high leverage positions. These positions include third base (3B), left field (LF), and center field (CF). Players at these positions contribute drastic amounts of value in both their batting, and fielding aspects of the game. These players tend to be overall strong players and can contribute at most levels of the game. The other position which stuck out was first base (1B), but first basemen rarely contribute anything meaningful on the defensive side of the ball. First basemen tend to have a more relaxed fielding responsibility but they tend to be some of the best hitters all around. First basemen are known to be phenomenal power hitters, and home runs and runs batted in are two of the most valuable contributors to value in terms of `fWAR`.

5.4 Impact of Player Attributes

Predicting a player's success based upon their draft information, demographics, and characteristics proved to be a more difficult task. While weight and height are important variables, we do not expect they have a phenomenal performance as the beauty of baseball is described by the randomness of the game.

5.5 Model Results

We were able to produce a model with a cross-validated accuracy of about 78.5%. By no means is this accuracy low, but top performing accuracy models will usually be in the mid 80s to higher in accuracy. The problem lies with predicting which players will be performing at a high level. The model was excellent at predicting the players who would perform poorly in their careers (`fWAR_class`: "Less than 5 fWAR"), but struggled at accurately predicting those who performed well during their major league careers.

The meaningful outcome from the model lies in the variable importance of the model constructed. It does not imply that all the variables are truly important, but there are a few which likely do have a strong impact on player success. There a 6 important variables we want to highlight from the model: - `debut_age` - `yrs_before_debut` - `pick_number` - `person_weight` - `person_height` - `draft_age`

Bibliography

Caporale, Tony, and Trevor C. Collier. 2013. "Scouts Versus Stats: The Impact of*Moneyball*on the Major League Baseball Draft." *Applied Economics* 45 (15): 1983–90. https://doi.org/10.1080/00036846.2011.641933.

Conforti, Christian M., Ryan L. Crotin, and Jordan Oseguera. 2022. "Major League Draft WARs: An Analysis of Wins Above Replacement in Player Selection." *Journal of Sports Analytics* 8 (1): 77–84. https://doi.org/10.3233/jsa-200586.

Crotin, Ryan L., Christian M. Conforti, David J. Szymanski, and Jordan Oseguera. 2023. "Anthropometric Evaluation of First Round Draft Selections in Major League Baseball." *Journal of Strength & Conditioning Research* 37 (8): 1609–15. https://doi.org/10.1519/jsc.0000000000004442.

Garmon, Christopher. 2012. "Major League Baseball's First Year Player Draft." *Journal of Sports Economics* 14 (5): 451–78. https://doi.org/10.1177/1527002511430229.

Staudohar, Paul D, Franklin Lowenthal, and Anthony K Lima. 2006. "The Evolution of Baseball's Amateur Draft." *NINE: A Journal of Baseball History and Culture* 15 (1): 27–44. https://doi.org/10.1353/nin.2006.0056.