# Projecting Professional Placement of Summer League Players

Patrick Mellady

## Objective

This report is concerned with the answer to the following question:

*Given season level data, we would like you to predict the professional outcomes of college players*

What follows is the solution to this question for certain players and how this information can be used for the purposes of player recruitment.

## Background

To answer this question, it is necessary to draw distinctions between potential professional outcomes for players. For the purposes of this analysis, professional outcomes for baseball players will have three categories: None, Minors, and Majors. These three categories indicate if a player has no professional placement, is a Minor League player, or is a Major League player, respectively.

Since the predictions will be based on statistical data, and pitchers and hitters have different recorded statistics, we must fit separate models for each of the two types of players. We will use the familiar recorded statistics, a portion of which are shown for hitters and pitchers below.

Some Familiar Batting Statistics

| Outcome | Name | Age | PA | AB | R | H | X2B | X3B | HR |
|---|---|---|---|---|---|---|---|---|---|
| None | Heiser, Grant | 22 | 63 | 381 | 326 | 49 | 85 | 18 | 1 |

Some Familiar Pitching Statistics

| Outcome | Name | Age | W | L | W.L. | ERA | RA9 | G | GS |
|---|---|---|---|---|---|---|---|---|---|
| None | Segel, Kaden | 20 | 3 | 2 | 0.6 | 4 | 5.25 | 24 | 0 |

Using the above statistics, predictions, which denote projected professional outcomes, will be fit for all players in the data set. The data have been scraped from Baseball-Reference.com and contain College career statistics for 651 hitters and 591 pitchers.

## Methodology

This problem can be thought of in terms of a multi-category ordered classification (i.e. the "None" category is a lower order than the "Minors" category, which is a lower order than the "Majors" category and the goals is to find into which category a player will fall). Classically, statisticians have developed methods for analyzing this kind of data by using what is called a *Proportional Odds* model. Intuitively, this works by using each

player's career statistics to estimate the probability of ending up in each of the three categories, and then make the final projection by whichever category has the highest probability. The benefits of the proportional odds model are that it can numerically quantify how influential each predictor is on professional placement and, since we have these numerical quantification, we can perform statistical tests to determine if certain information is influential at all.

An alternative approach to this type of problem is to use classification trees or random forests. These methods have the benefit of both being easier to understand and not relying on assumptions about the structure of the data and can thus better model complex relationships. However, with these tree based methods, there is no ability to quantify exactly how influential each predictor is on professional placement, which is often desirable.

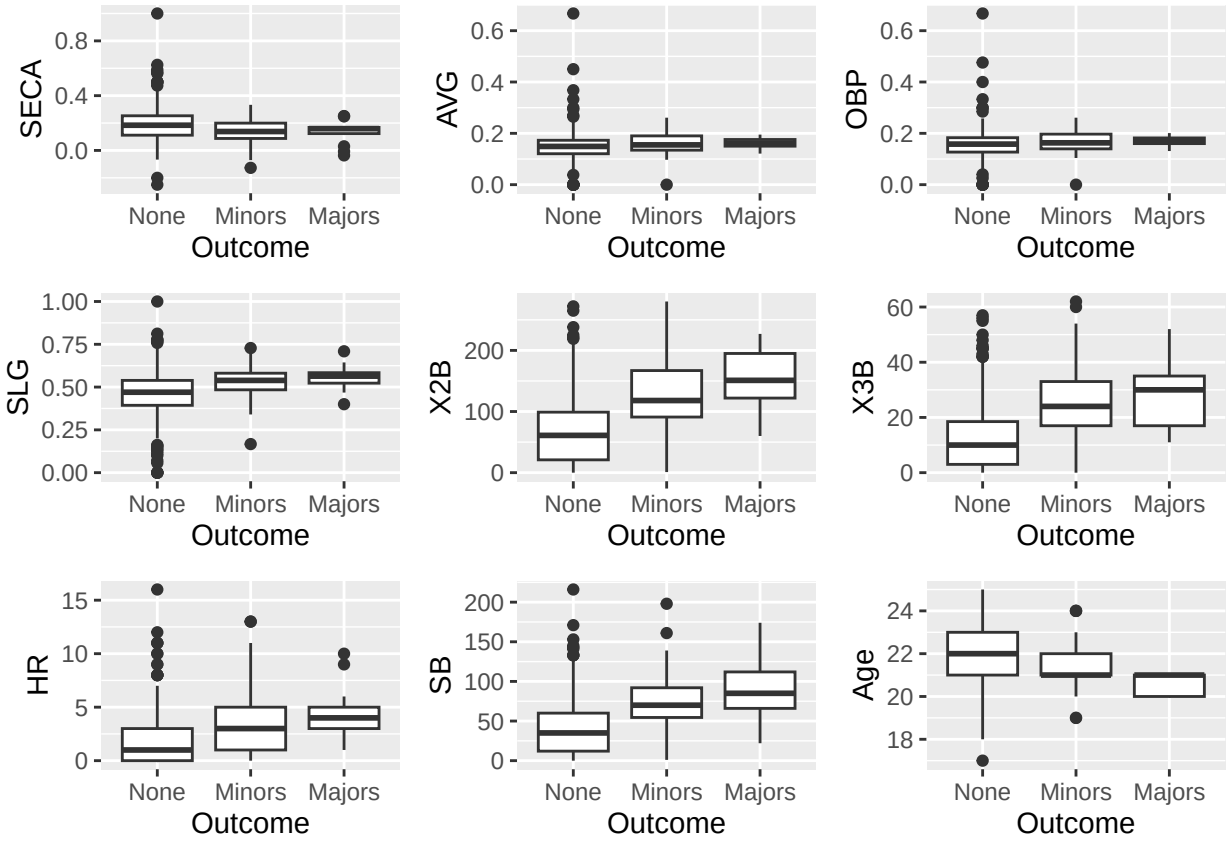
In this report, both the proportional odds and random forests models are fit to the data.

# Results

## Batters

### Exploratory Analysis

To see which statistics are the most related to a batter's professional outcome, below shows some exploratory data analysis via data visualization.




The above suggests that player placement is more affected by counting-based statistics than rate-based statistics, which may be because of limitations on playing time for some players leading to outliers. Additionally, see that being 21-years-old in your final NCAA season is associated with playing at the professional level, while being older or younger is associated with not being drafted. With these insights, counting-based statistics will be used when fitting the parametric model.

**The Proportional Odds Model**

The proportional odds model has been fit to the batters data, a random sample of players, their current placements, and their predicted placements is shown below.

Players Actual vs Predicted Placements (PO Model)

| Name | Current.Placement | Projected.Placement |
|---|---|---|
| Saum, Charlie | None | None |
| Legg, Jimmy | None | Minors |
| Atwood, Andy | Minors | Minors |
| Suddleson, Jake | Minors | Majors |
| Spohn, Harrison | Minors | Minors |
| Oakley, Nick | None | None |
| Oyama, Joichiro | None | None |

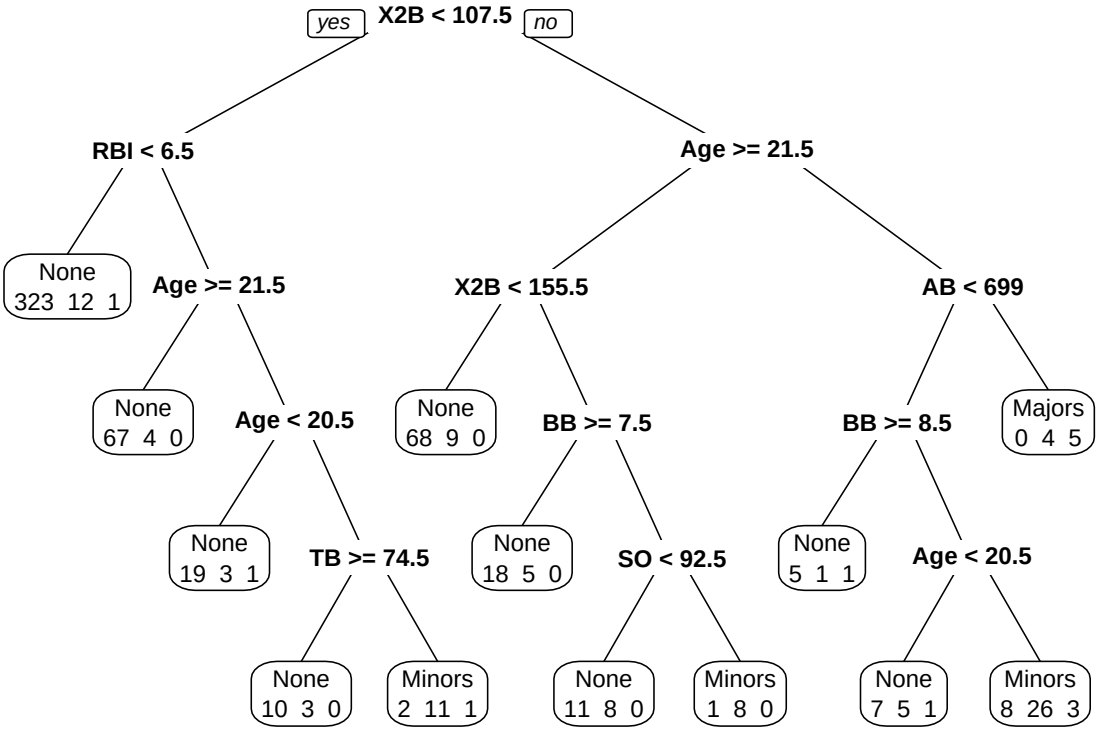
The following table summarizes the accuracy of this method. This table shows what percent of players from each of the observed outcomes were predicted in each category. For example, of the players who have an observed outcome of "None", 96.8% were correctly predicted as none, 3.2% were incorrectly predicted as "Minors" and 0% were incorrectly predicted as "Majors". The entries along the main diagonal indicate that the model is best at predicting players whose outcome will be "None", and gets worse as the outcome level increases. Lastly, the proportional odds model has an overall accuracy of 91.3%.

Accuracy of PO Model for Batters

| | Pred. None | Pred. Minors | Pred. Majors |
|---|---|---|---|
| Obs. None | 96.8% | 3.2% | 0% |
| Obs. Minors | 27.6% | 69% | 3.4% |
| Obs. Majors | 22.2% | 22.2% | 55.6% |

**Random Forests**

The non-parametric model is a random forest. To visualize how this method works, the outcome a classification tree is shown below:




This graphic shows a series of yes/no questions that will result in a professional placement for every player. If a player has 107 or fewer doubles and 6 or fewer RBIs, that player will be classified as "None". A random forest works by fitting multiple trees to different samples of the data and averaging the predictions. Here is a sample of players, their placements, and projections produced by the random forest method:

Players Actual vs Predicted Placements (Random Forest)

| Name | Current.Placement | Projected.Placement |
|---|---|---|
| Guerrero, A.J. | None | Minors |
| Shimao, Tate | None | None |
| Matthiessen, Will | Minors | Minors |
| Rutschman, Adley | Majors | Majors |
| Reyes, Ripken | Minors | Minors |
| Cano, Willie | None | None |
| Saunders, Ty | None | None |
| Gibson, Hunter | None | None |
| Ostrander, Garret | None | None |
| Paganelli, Adam | None | None |
| Fullerton, Maxim | None | None |

Similar to the parametric model, the table below summarizes the accuracy of this method. Random forests work significantly better than the proportional odds model, suggesting that the assumptions of the proportional odds model may not be satisfied by the data. The main diagonal shows that the random forest has some

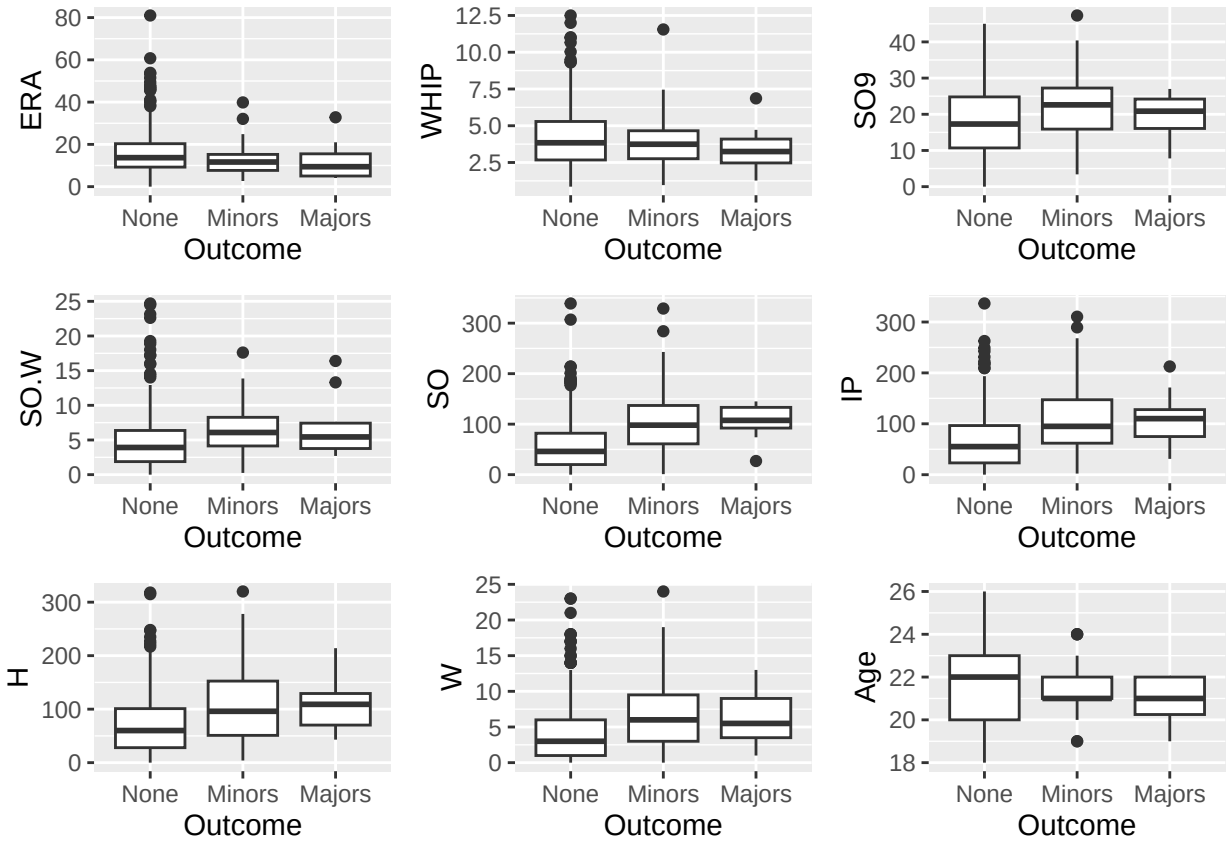
difficulty separating players who will have no professional placement from players who will be placed in the Minors, but this error is still an improvement over the proportional odds model. The random forest has an overall accuracy of 99%.

Accuracy of Random Forest for Batters

| | Pred. None | Pred. Minors | Pred. Majors |
|---|---|---|---|
| Obs. None | 99.8% | 0.2% | 0% |
| Obs. Minors | 4.6% | 95.4% | 0% |
| Obs. Majors | 0% | 0% | 100% |

## Pitchers

What follows is a repetition of the methods above applied to the data set of pitchers. ### Exploratory Analysis To see which statistics are the most related to a pitcher's professional outcome, below shows some exploratory data analysis via data visualization.




Similar to what was seen with the batters, rate-based statistics do not provide clear separation between the professional placements while counting-based statistics do. The trend relating a players age in their final college season and placement is also seen here with the pitcher's data. With the insights from above, the two models are fit and analyzed below in the same manner as was done for the batters.

### The Proportional Odds Model

A proportional odds model has been fit to the pitcher's data using the counting statistics, as indicated from the exploratory analysis above. A random selection of players, their outcomes, and projections is produced

below:

Players Actual vs Predicted Placements (PO Model)

| Name | Current.Placement | Projected.Placement |
|---|---|---|
| Haarer, Toby | None | None |
| Lucas, Easton | Majors | Minors |
| Giroux, Alex | None | None |
| Guarin, Marcus | None | None |
| Engman, Jared | None | None |
| Peery, Brock | None | None |
| Sanchez, Corey | None | None |
| Stumbo, Peyton | Minors | Minors |
| Walters, Carson | None | None |

The table below shows the accuracy of this method.

Accuracy of PO Model for Pitchers

| | Pred. None | Pred. Minors | Pred. Majors |
|---|---|---|---|
| Obs. None | 95.8% | 4.2% | 0% |
| Obs. Minors | 34.9% | 65.1% | 0% |
| Obs. Majors | 14.3% | 57.1% | 28.6% |

Again, this table shows that the proportional odds model is worse at predicting professional placements as the level of play increases. The proportional odds model for pitchers has an overall accuracy of 88.8%.

**Random Forests**

A random forest was fit to the pitchers data and a random sample of players, their observed placements, and projected placements is produced below.

Players Actual vs Predicted Placements (Random Forest)

| Name | Current.Placement | Projected.Placement |
|---|---|---|
| Pinedo, Andrew | None | None |
| Prizina, Jake | Minors | Minors |
| Tongue, David | None | None |
| Barry, Shea | Minors | Minors |
| Horak, Jordan | None | None |
| Pilchard, Cade | None | None |
| Spoljaric, Garner | None | None |
| Mathews, Quinn | Minors | Minors |

The following table summarizes the accuracy of this method:

Accuracy of Random Forests for Pitchers

| | Pred. None | Pred. Minors | Pred. Majors |
|---|---|---|---|
| Obs. None | 100% | 0% | 0% |
| Obs. Minors | 7% | 93% | 0% |
| Obs. Majors | 0% | 0% | 100% |

Similar to what was seen for the batters, the random forest method does a better job at predicting the professional placement of pitchers. The overall accuracy of the random forest for pitchers is 99.3%.

## Conclusion

Although both approaches make errors by under-placing players (projecting them to end at a lower level than observed), random forests seem to model the complex relationship of the data better than the proportional odds model. For the proportional odds model, the larger error problem should be resolved by the introduction of more data, since, due to the age of the data, there are not many Major League appearances from this set of players. As time progresses and the data is recollected, the model can be refitted and retested to create more accurate projections.

While waiting for the recollection of data, it is recommended to use the random forest method for the projection of player outcomes as a basis for player recruitment. This method is more accurate and will provide more insight into which player should be acquired to increase revenue when compared to the proportional odds model.