Silver Slugger Award

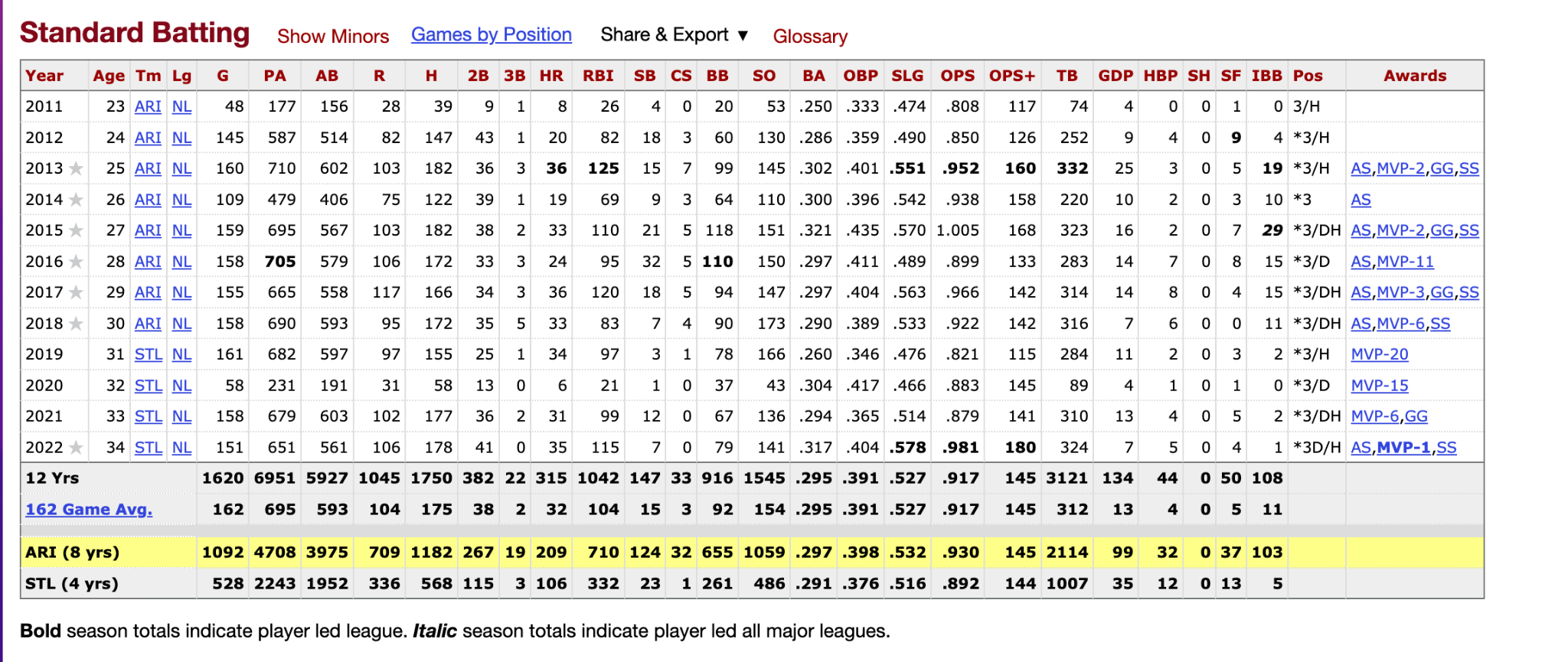
Executive Summary

## presented by

TJ Hamilton

## Data Preprocessing

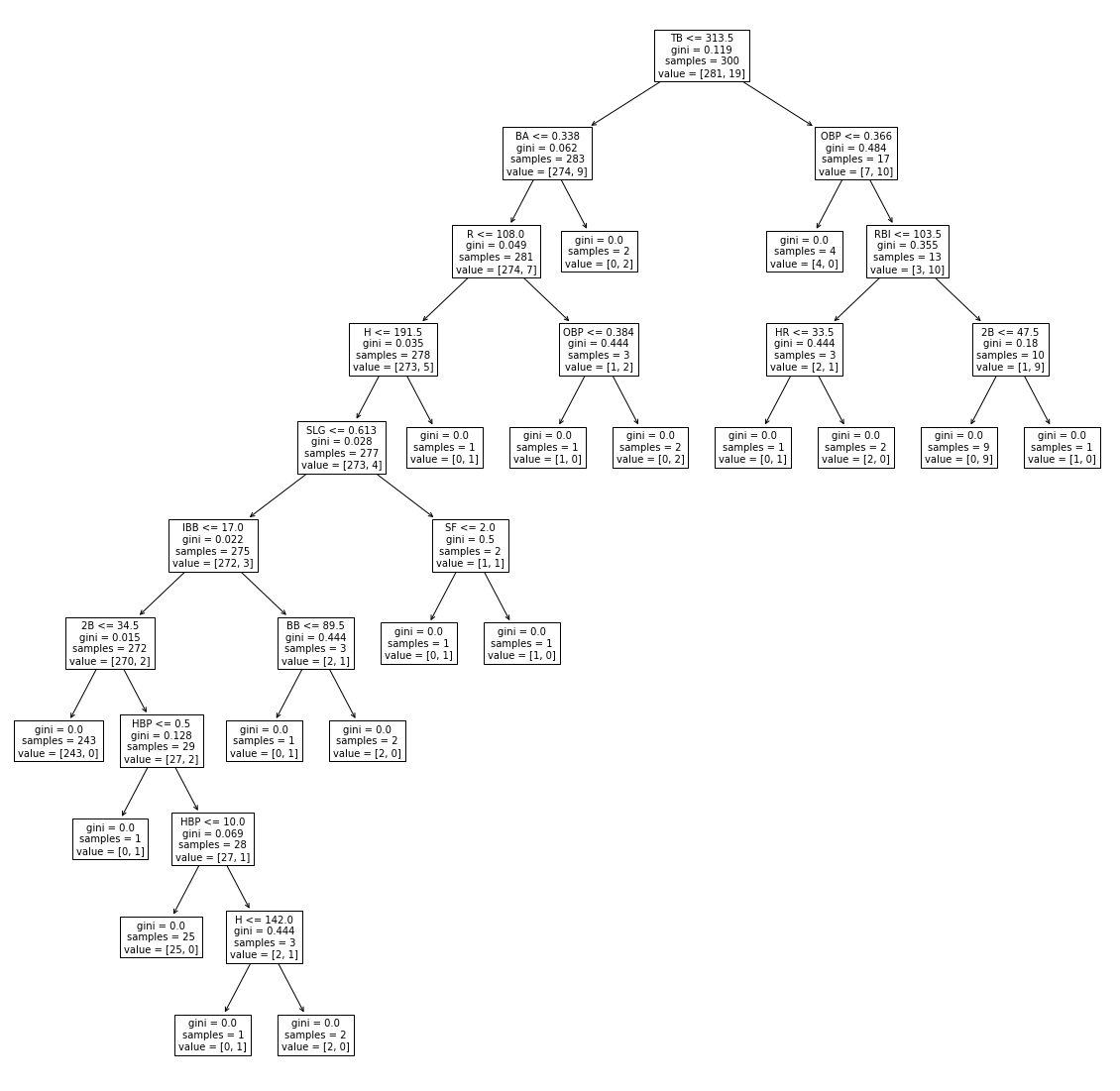
The website <https://www.baseball-reference.com/> contains data about every baseball player from every season since 1876. The following image is an example of the data I collected from the following website.



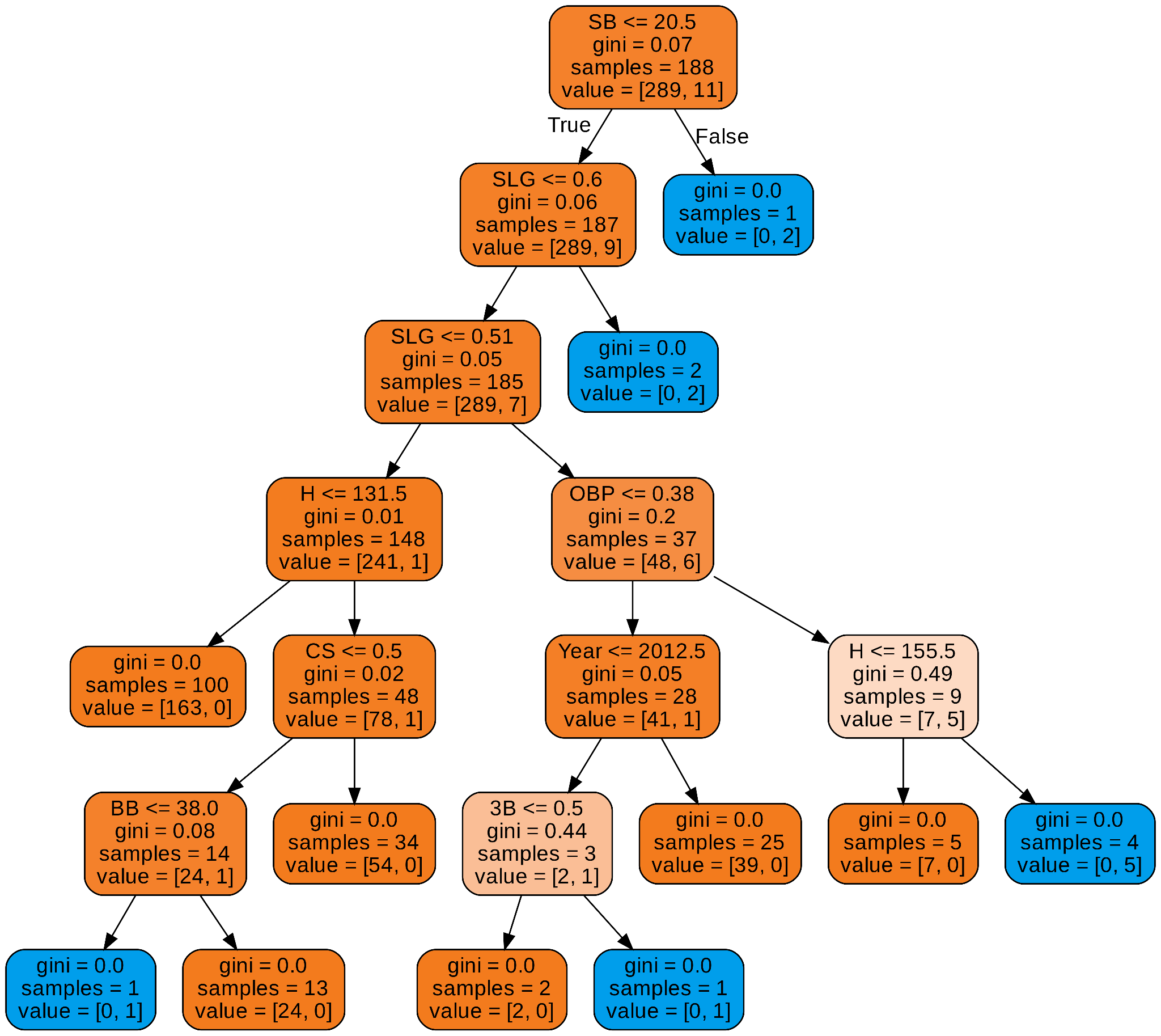
From this data I took every column except the OPS+ column. The OPS+ stat is basically the same as the OPS stat but the only difference is that it is adjusted for the specific player's home ballpark. I collected data from players that play First Base, and from the 2012 season up to this last season.

## Selection, training, and use of a machine learning model

Since I didn’t collect a lot of data, I took this last season out of the original data and used it to test how well the model does. Since I am trying to predict who won the silver slugger award, I decided to use a decision tree model. I tried decision tree classifier and random forest classifier models. For the decision tree model had the following structure:



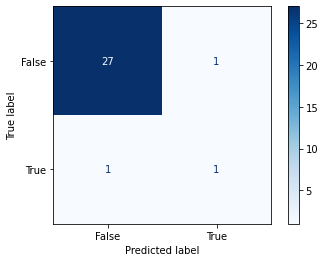
I find this model super interesting. It took almost every stat that was in the data set to classify if the athlete won the award. The random forest model had the following structure:

​​

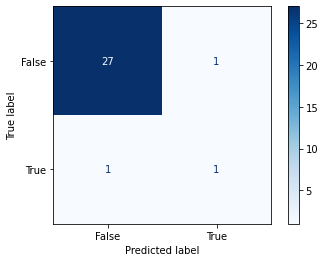
I like this model because it is a shorter tree than the model before.

## Interpretation of results (for example, using proper metrics, understanding limitations)

To see how well each model performs I created a confusion matrix for each specific model. This is the confusion matrix for the decision tree:



From this It tells me that the model was able to predict one of the winners, but predicted one wrong winner. The same thing happened with the random forest model:



## Communication of results

At this point in time, I would not use this model to predict which athlete would win the silver slugger award. This is because it was not able to pick up on some key information about the silver slugger award. For example, both models predicted two National League players to win the award this last season. It is good that the model learned that only two people can win the award in a given season, but it is only awarded to one athlete for each league. Both models predicted that the two winners of the award would be in the National League. 

The player that the model picked to win the award, but didn’t actually win it was declared to be one of the finalists this season for the National League. I think maybe even using a different model and using the data from 1980 (when the award first appeared) to today, there is big potential for it to work.

## Ethical implications

As of right now, the silver slugger award is given based on stats, and is chosen by a committee. The same thing for every end of season award. With the creation of models that could predict which player should win the award, there wouldn’t need to be very much debate in the meetings about which player should win the award. I remember when Paul Goldschmidt was on the Arizona Diamondbacks, he had incredible seasons, but wouldn’t be recognized for it because he was on a “low budget” team. Now with these models, the politics about which team the athlete is on would not have a huge impact on which player should win the award.

## Google Collab File

<https://colab.research.google.com/gist/Thamilton76/e1d32a4452ae24a5763130213ecfc689/silver-slugger.ipynb>