Predicting MLB Career Acolades from Rookie Seasons

Ashling Scott

May 4th, 2024

Introduction

This project is an effort to predict future career accolades for MLB players, using only the data from their first season in the league (their rookie season). Accomplished MLB players may play for a decade or more, and their first season is often nothing special. However, if front offices could figure out which players will accel from the start, they can acquire talented players for cheap, and ensure their team will have a successful player for the future.

However, this is incredibly difficult. Baseball is known as one of the most variable sports, which is part of why they have embraced analytics more than any other, trying to sort through the noise to find insights. Many players can have great first seasons and then fall of the map, while plenty of players may have been nothing special from the start but turned into legendary players. Anyone who could figure out career trajectory from only one season of data would be making a fortune somewhere in a team front office. So we must set our expectations accordingly; we will find no highly predictive models in this endeavour, we will simply do the best we can to make predictions.

The dataset was patched together from multiple datasets on MLB players, which required a significant effort of wrangling and pruning. The largest source of the dataset comes from Neil Paine on github: https://github.com/Neil-Paine-1/MLB-WAR-data-historical/tree/master. This dataset contains a huge amount of advanced analytic information, including per-162 rates (normalizing statistics out to 162 games, the length of a season), and WAR. WAR is an attempt at an all-encompassing value stat in baseball. It stands for Wins Above Replacement, and attempts to quantify how many extra wins a player nets their team when compared to a replacement level player, a player that could be acquired with little effort. There are many different types of WAR, as it has several different formulas and also can be calculated with certain restrictions. One of the main types I'll be looking at is JEFFBAGWELL WAR, which is a composite of two other main types of WAR known as bWAR and fWAR.

As target metrics, I had to find information on what players were in the Hall of Fame, and what players were selected as All-Stars. The Hall of Fame is a post-career accolade, reserved for only the greatest players of all time. An All-Star selection involves being selected for the yearly All-Star game, and is an accolade that a player can receive multiple times during their career. I acquired this data from Sean Lahman on Kaggle: https://www.kaggle.com/datasets/seanlahman/the-history-of-baseball. I stitched this hall of fame and allstar information together to get our full dataset, and then narrowed the data down to rookie seasons only.

Statistical learning strategies and methods

Our tool of choice for this analysis will be logistic regression, using target variables of hall_of_fame and allstars. This allows us to make binomial models and make predictions based on them. I've divided the data into two types: Hitters and Pitchers. Hitters and pitchers each have different statistics used to track their performance, and also tend to have different career trajectories. They're different enough that they need to be seperated for thorough analysis.

We'll start by looking at correlation matrices. We can check correlations for both hitters and pitchers for all-star selections as well as hall of fame inductions. These numbers should clue us into what the most predictive features are, and will guide building our models.

```
##                    value
## hall_of_fame 1.00000000
## allstars     0.24462033
## WAR162       0.18147218
## bwar162      0.18147118
## pa           0.13144368
## bat162       0.12921098
## rep162       0.11969794
## pct_PT       0.11799506
## g_bat        0.11164205
## fld162       0.04982768
```

This is the correlation data for hitters and hall_of_fame. Let's break down what these figures actually mean. Unless otherwise mentioned, all these stats are taken from the player's rookie season alone.

hall_of_fame: Our target variable for this section of the analysis, and indicates whether a player was placed in the Hall of Fame or not.
allstars: Whether or not a player was ever selected as an All-Star in their career.
WAR162: JEFFBAGWELL WAR per 162 games.
bwar162: This is batting war 162, which is WAR that comes from batting specifically.
pa: Plate appearances, which is how often a hitter had a chance to bat.
bat162: Batting Runs Above Average per 162 games.
rep162: Replacement runs per 162 games.
pct_PT: The player's share of the team's total playing time.
g_bat: Games that the player batted in.

When building our model, we want to favor predictors that have high correlations, but several of these predictors are unsuitable. Obviously we can't use hall_of_fame since it's our target variable, and we also don't want to use all-stars because its data that's usually only available in hindsight. We want to make predictions based on the players rookie season, not look ahead to see if they were ever an all-star, and players are rarely all-stars during their rookie season. Many of the WAR stats are very similar, and provide no additional predictive power, so we will generally want to just choose the strongest one and leave the others out.

Our final logistic model is: hall_of_fame ~ WAR162 + pa + bat162 + rep162 + pct_PT + g_bat

Now let's take a look at the correlations for all-star selection.

```
##                   value
## allstars     1.00000000
## hall_of_fame 0.24462033
## pa           0.19208502
## WAR162       0.19178563
## bwar162      0.19178194
## rep162       0.17809641
## g_bat        0.17774864
## pct_PT       0.17251738
## year_ID      0.12128608
## bat162       0.08290068
```

The most interesting thing here is that plate appearances outdoes the WAR stats as a predictor. This suggests that being good enough in your rookie season to get a lot of at-bats is more important than your numerical performance for becoming an all-star at some point in your career. This indicates that maybe getting a lot of playing time in your rookie season may be an indication of all-star level talent, but you would need impressive WAR stats to make the Hall of Fame, which is a more difficult goal.

Let's check out some pitcher correlations now:

```
##                   value
## hall_of_fame 1.00000000
## allstars     0.21445401
## fg_pwar162   0.06825467
## pwar162      0.06692951
## WAR162       0.06593567
## br_pwar162   0.06281690
## ra9_pwar162  0.06076792
## Kpct_plus    0.05844145
## pos162       0.05733965
## def162       0.05699953
## innings      0.05495750
## pct_PT       0.05263963
## K9_plus      0.04597363
## starts       0.04538947
## pa           0.04409054
```

To explain the new stats here:
Kpct_plus: Strikeout rate compared to league average
innings: Total number of innings pitched
K9_plus: Strikeouts per 9 innings compared to league average
starts: How many games the pitcher started

The pitcher hall of fame correlations are really weak across the board. The highest usable value being only 0.06 paints a grim picture for the predictive power of any potential models. What's more, most of these are just different forms of WAR, of which we are only going to use one. We can expect this model to be the weakest so far.

```
##                   value
## allstars     1.0000000
## fg_pwar162   0.2710768
## WAR162       0.2593892
## pwar162      0.2573738
## ra9_pwar162  0.2430395
## br_pwar162   0.2251121
## innings      0.2157558
## hall_of_fame 0.2144540
## starts       0.2130325
## pct_PT       0.2084852
## def162       0.1752769
## pos162       0.1747577
## Kpct_plus    0.1395074
## pa           0.1251980
## g_pitch      0.1093107
```

In contrast to hall of fame, the pitcher correlations to all-star appearances are very strong, perhaps speaking to the difficulty of finding truly transcendant players versus finding those that can be good for one season. Maybe the most interesting thing about this list is how low hall_of_fame is. I expected allstars and hall_of_fame to be the strongest correlation on every correlation matrix, but here it's well below all the WAR stats. This may be due to the lower career duration that pitchers often have, due to injury or wear and tear, suggesting that pitchers can have season of greatness that don't translate as often to legendary careers.

In order to select our models, we'll take these correlation matrices as a guide and perform Forward Stepwise-Selection. The resulting models are as follows:

Hitter HoF: hall_of_fame ~ WAR162 + pa + bat162 + rep162 + pct_PT + g_bat
Hitter Allstar: allstars ~ pa + WAR162 + rep162 + g_bat + pct_PT + year_ID
Pitcher HoF: hall_of_fame ~ fg_pwar162 + Kpct_plus + pos162 + def162 + innings + starts
Pitcher Allstar: allstars ~ fg_pwar162 + innings + starts + pct_PT + def162 + pos162 + Kpct_plus

As a final note, the distribution of our target variables is extremely imbalanced. Only around 1% of players make it into the hall of fame, and less than 10% of players ever get an all-star selection. This is going to cause problems in our models, and we'll need to combat these problems. Our method of choice will be class weights, giving stronger weight to the sparser side of the logistic regression, in order to encourage the model to make bolder predictions in that direction.

Predictive analysis and results

Hitters

With our models selected, we run them through 10-fold cross validation using the caret library. We use this to make predictions so that we can check the predictive strength of our model on the test data. After the test data from each fold is compiled, we can create confusion matrices, which show us not only which predictions are correct, but whether each prediction is a True Positive, False Positive, True Negative, or False Negative.

Table 1: Hitter - Hall of Fame

	Negative	Positive
Negative	10788	1
Positive	162	6

Our first confusion matrix shows some serious problems. While the accuracy rate is high at 98.51%, the table still barely performs better than always guessing Negative. This is because the dataset is so imbalanced: Almost 99% of the data from our dataset is Negative. It's very difficult for player's to get into the Hall of Fame, as you might expect. Logistic regression has trouble dealing with this imbalance, since by default it's only looking at overall accuracy, instead of the accuracy of False Positives or False Negatives. Here we can see a shockingly bad rate of False Negatives, as 162 out of 168 true Positive values have been classified wrong.

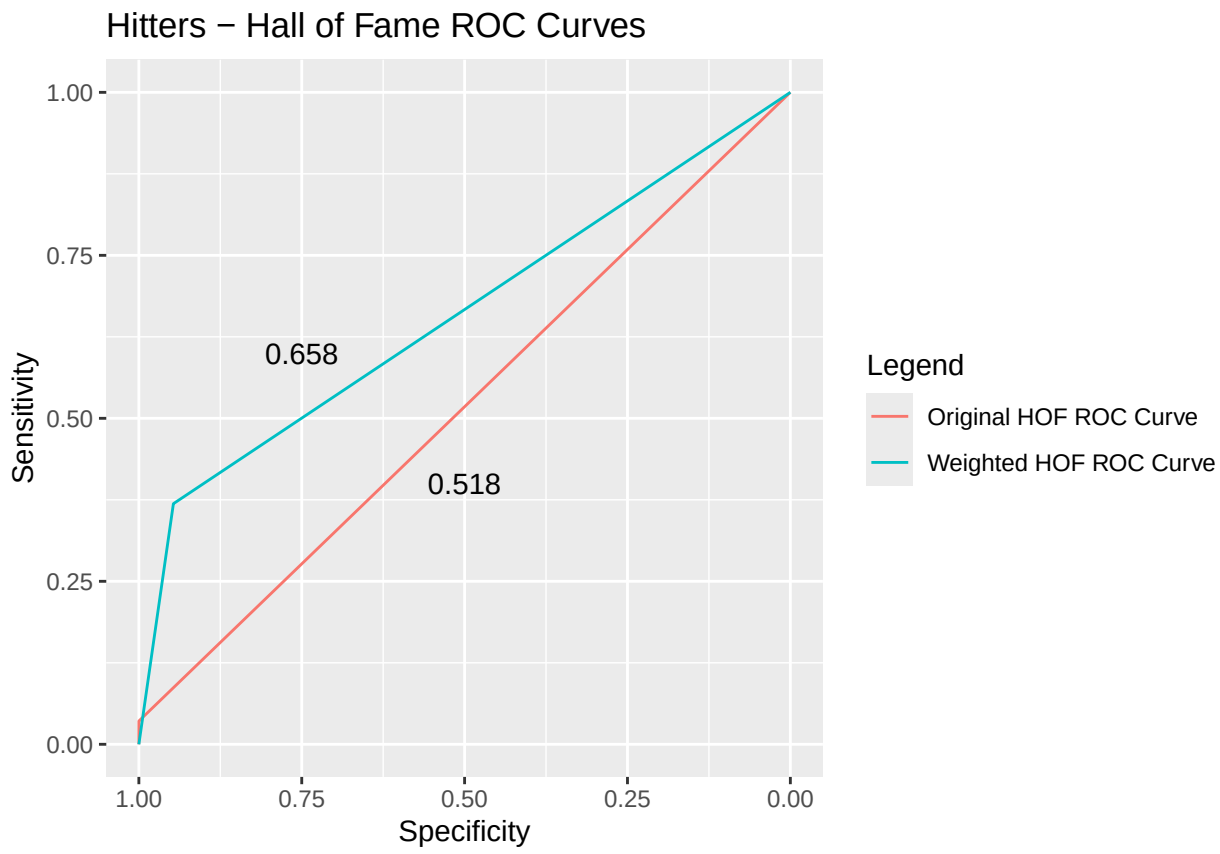
One way we can address this issue is by weighting our results. In order to convince our model not to prioritize accuracy rate over everything, we can weight the different classes differently, accepting a much higher rate of False Positives in order to find more True Positives. After some fiddling with the data, I found the best weight for this situation was 30:1, in other words we are weighing Positive results at 30 times the value of Negative results, in terms of finding the right balance. After changing to weighted values, the new confusion matrix is as follows:

Table 2: Weighted Hitter - Hall of Fame

	Negative	Positive
Negative	10217	572
Positive	106	62

Here we see a more balanced distribution, with a whole 62 True Positives. There's still a lot of False Positives and False Negatives, but the model is now actually giving us some Positive guesses, so it's a good tradeoff.

We can get an idea of how effective this weighting was by looking at ROC curves of the original model versus the weighted model.



Here we see ROC curves of the two models, one original and one weighted. The ROC curve of the original model shows just how weak it is, due to the immense imbalance. The area under the curve is barely higher than 0.5. The weighted model manages a more respectable AUC, at 0.658. This wouldn't be a strong value in most scenarios, but given the difficulty of the problem and the severe imbalance of the data, it's a reasonable score.

Moving onto allstars as the target variable, we will need to take a fresh look at the confusion matrices.

Table 3: Hitter - Allstars

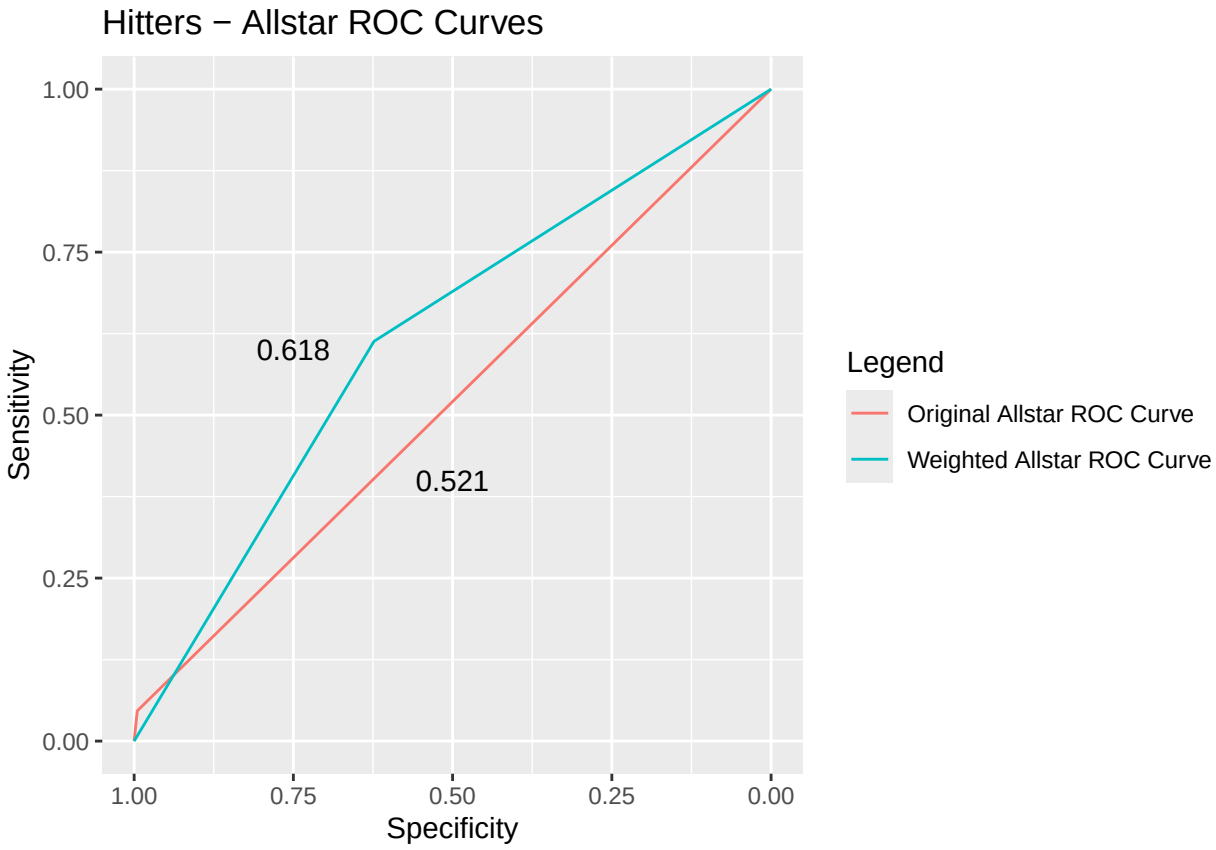
	Negative	Positive
Negative	9637	48
Positive	1213	59

We see similar problems here. While the rate of False Negatives isn't quite as bad as the original HoF confusion matrix, its still a huge issue. We will once again apply weight to allstar selections, not quite as heavy this time though, only 10:1 ratio.

Table 4: Weighted Hitter - Allstars

	Negative	Positive
Negative	10217	572
Positive	106	62

This is a much more acceptable rate of Positives, getting over 1/3rd of them correctly. Let's take a look at the ROC curves.



The improvement isn't as large here, given that the Original Allstar ROC Curve is slightly better than the Original HOF Curve, while simultaneously the Weighted Allstar curve has a lower AUC value.

Pitchers

Table 5: Pitcher - Hall of Fame

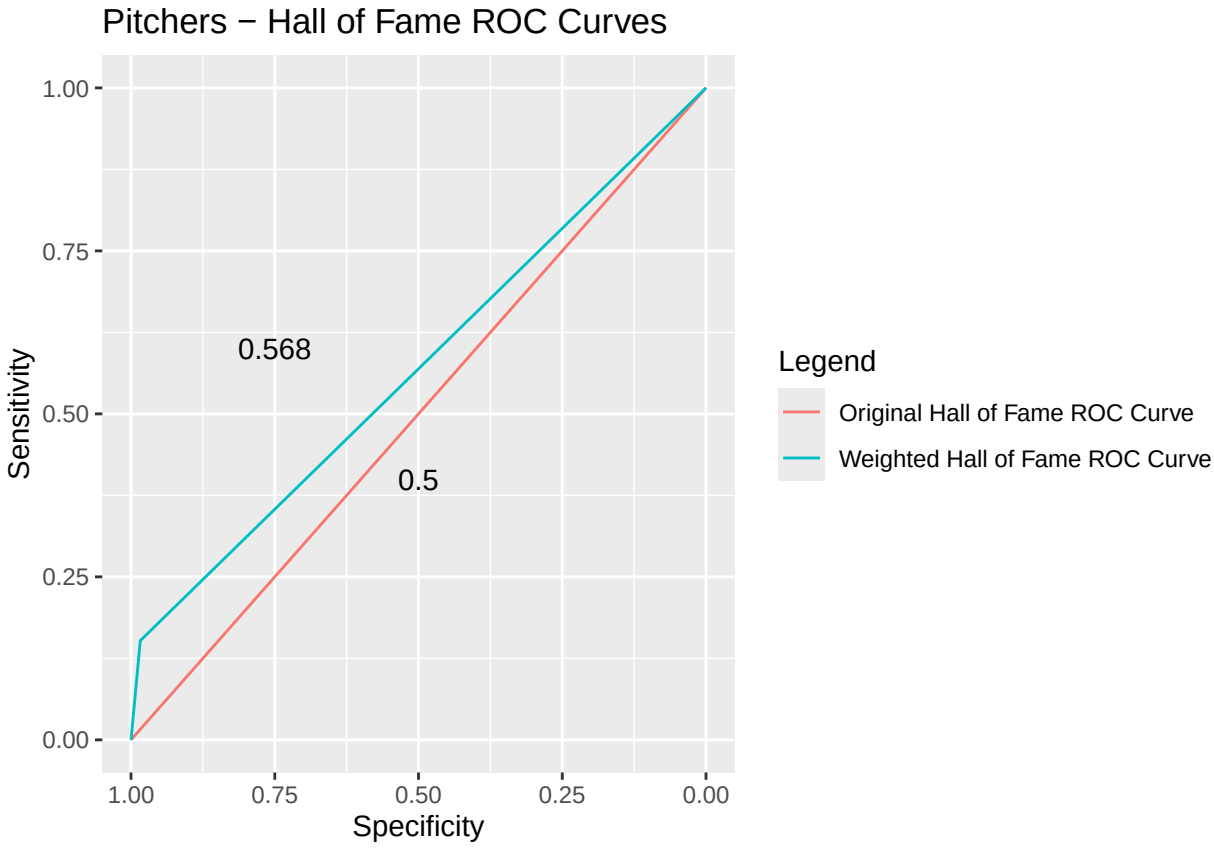
	Negative	Positive
Negative	7618	0
Positive	46	0

As we feared when looking at the super low correlation rates for hall_of_fame pitchers, this model has failed completely. It didn't guess that a single player would be a hall of famer, simplying assuming that every single player would fail to reach it. I'm not sure even weighting will be able to salvage this situation.

Table 6: Weighted Pitcher - Hall of Fame

	Negative	Positive
Negative	7495	123
Positive	39	7

After a heavy weighting of 30, we still have a pretty poor model. Let's check the ROC curves.



We only manage an AUC of 0.568, which is a marginal improvement over the flat 0.5 original rate that guessed the same result every time. Hopefully the allstars data is more workable, with its higher correlation rates.

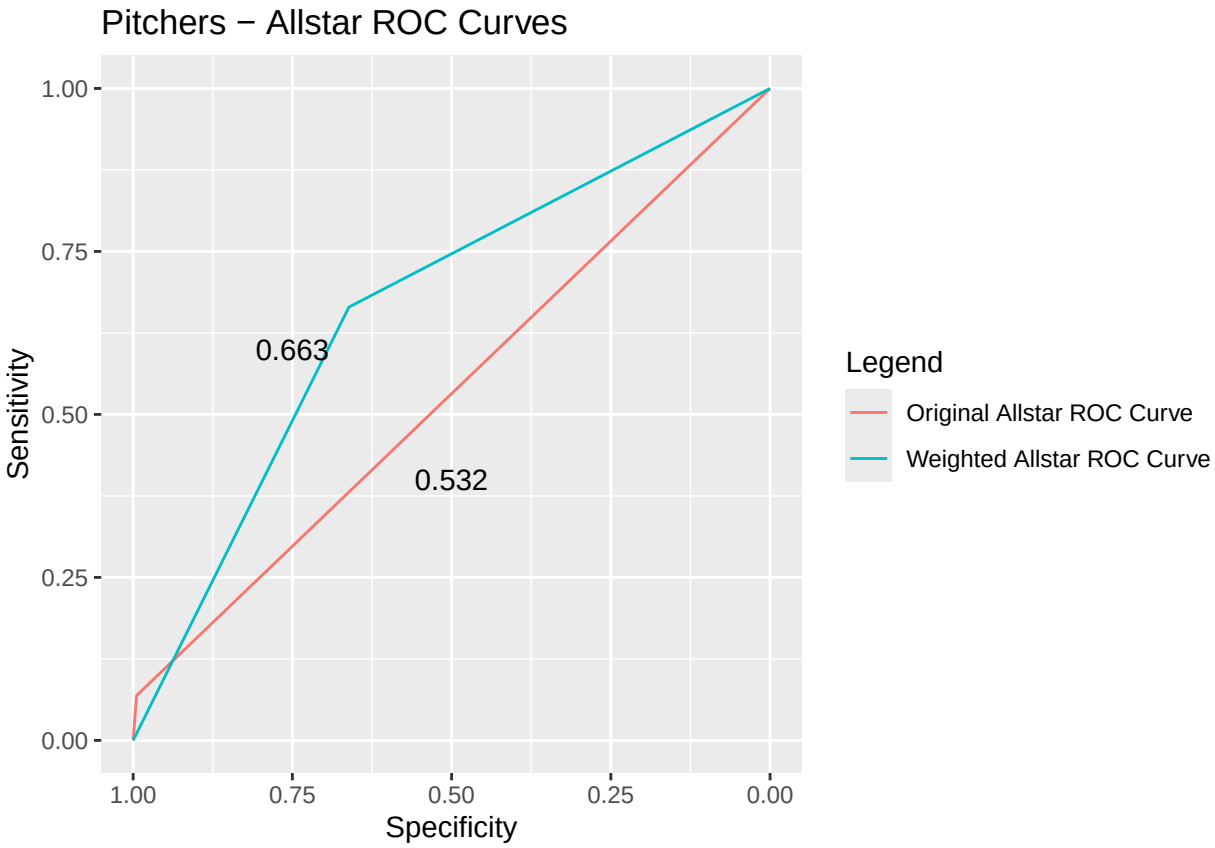
Table 7: Pitcher - All-Star

	Negative	Positive
Negative	6813	37
Positive	758	56

Table 8: Weighted Pitcher - All Star

	Negative	Positive
Negative	4529	2321
Positive	273	541

These matrices are a bit more agreeable, especially after adding a weight of 10 to allstars.



This is the best ROC curve of all the ones we've looked at, with an AUC of 0.663. It's interesting to me how it was much easier to predict allstar selection than hall of fame induction for pitchers, while it was the complete opposite for hitters, which made it an excellent idea to look at the categories separately.

CONCLUSION

Our models performed about as successfully as we expected, given the extreme difficulty of the task. Still, it's mostly guesswork, and the majority of our predictions for players to receive accolades will still be wrong no matter what we do. Hall of Fame turned out to be slightly easier to predict for hitters, while all-star status was easier to predict for pitchers, the reasons for which I have speculated on throughout this analysis.

Our projects scope was total, comprising of nearly every MLB player of the last 100 years (some players had critical missing data). Its arguable that having the scope be all of MLB history was a mistake; perhaps the formula for getting allstar or hall of fame selections has greatly changed over the many decades and older data is poisoning our modern insights. Using only more modern data might also make the data more generalizable, since the factors that existed in older time periods may no longer have any consideration to the modern era.

It also could be the case that a player's first season might not be the best indicator of future success. Players generally have several years on their low-paying rookie contract before they are eligible to be resigned. A front office might be more interested in data from the first several years instead, since those are the years where they will be able to get data before having to pay a player significantly more money. It's very likely the case that several years of data produces a much stronger model to predict which players will accel.

To close things out, let's list out all our 2022 (the most recent year in our dataset) rookies that our model has predicted will be in the hall of fame. We'll check back in 50 years to see how well things panned out.

```
## # A tibble: 15 x 5
## # Groups:   key_bbref [15]
##    player_name          age key_bbref year_ID team_ID
##    <chr>              <int> <chr>       <int> <chr>
##  1 Brendan Donovan       25 donovbr01    2022 STL
##  2 Oscar Gonzalez        24 gonzaos01    2022 CLE
##  3 Riley Greene          21 greenri03    2022 DET
##  4 Michael Harris        21 harrimi04    2022 ATL
##  5 Steven Kwan           24 kwanst01     2022 CLE
##  6 MJ Melendez           23 melenmj01    2022 KCR
##  7 Joey Meneses          30 menesjo01    2022 WSN
##  8 Jose Miranda          24 miranjo01    2022 MIN
##  9 Christopher Morel     23 morelch01    2022 CHC
## 10 Vinnie Pasquantino    24 pasquvi01    2022 KCR
## 11 Jeremy Pena           24 penaje02     2022 HOU
## 12 Julio Rodriguez       21 rodriju01    2022 SEA
## 13 Adley Rutschman       24 rutscad01    2022 BAL
## 14 Seiya Suzuki          27 suzukse01    2022 CHC
## 15 Bobby Witt            22 wittbo02     2022 KCR
```