What does it take to enter the Hall of Fame?

Analyzing extensive data through machine learning

Authored by:

Sander Bathke, Alperen Bayar

One of the greatest honors a baseball player can receive is to be inducted into the Hall of Fame, an elaborate multi-step voting process that results in 3-5 additions per year. We wanted to explore what light machine learning can shed on this process, boiling down the complex opinions of human experts to a statistical “hall of fame” of the strongest predictors.

We used the Lahman database as our source, a large database created and maintained by a data journalist tracking all Major League games since 1871 with annual updates. We merged different sheets to summarize every player’s performance into 130 data points including total, average, best, and worst performances in various metrics.

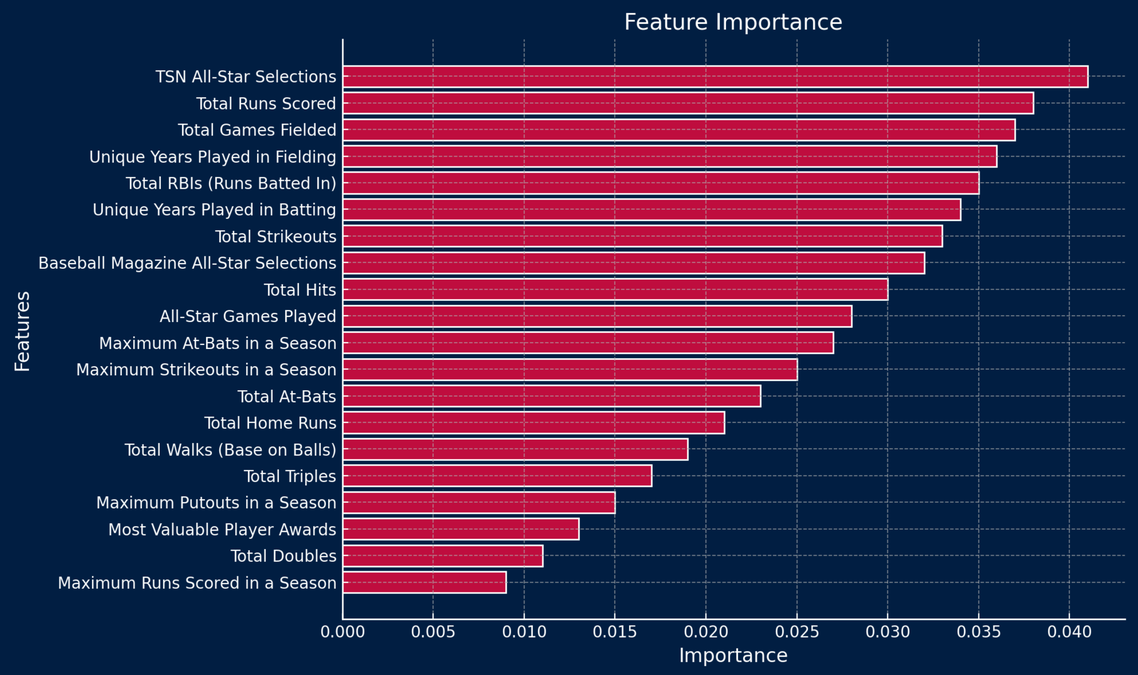
We tried a total of three algorithms to glean information from the data:

	Random Forest Classifier	Naive Bayers Classifier	Custom Neural Network
Accuracy	88%	80%	80%
Precision (not HoF)	91%	85%	80%
Precision (HoF)	71%	49%	0%
Precision (Wgt. Avg.)	87%	78%	64%
Recall (not HoF)	93%	90%	100%
Recall (HoF)	65%	37%	0%
Recall (Wgt. Avg.)	88%	80%	80%
F1-Score (not HoF)	92%	88%	89%
F1-Score (HoF)	68%	42%	0%
F1-Score (Wgt.Avg.)	88%	78%	71%

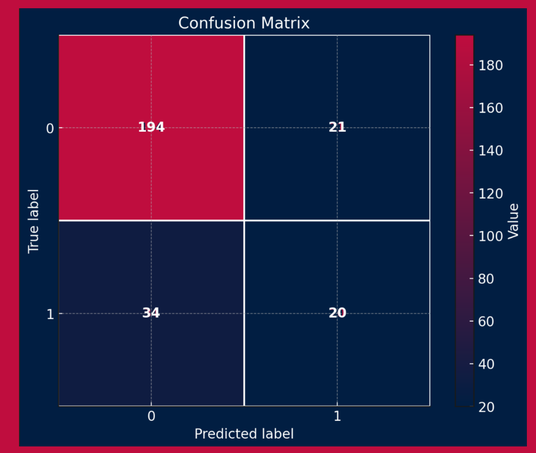
As the Random Forest Classifier proved the most powerful predictor, we derived our further information from it alone.

To get into the US National **Baseball Hall of Fame**, you must play for a **long time** and have **a season** with **exceptional performance.**

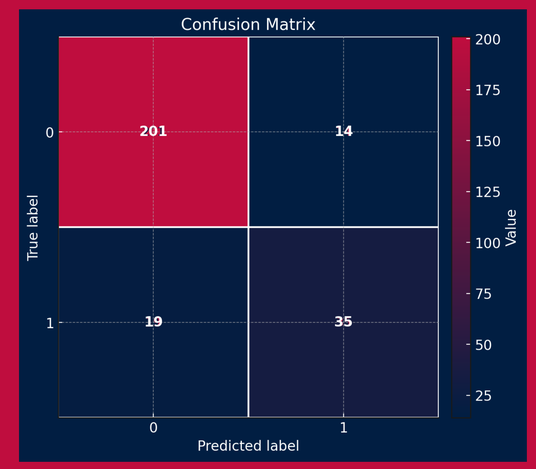
The **Top 20 most important** statistics are:



Confusion Matrix: Naïve Bayers



Confusion Matrix: Random Forest



Non-Players in the Hall of Fame