Baseball Trends: Predicting Future Hall of Famers

Aurin Chakravarty, Rohan Tilva, Ashish Phal
AMS 550.436, Johns Hopkins University

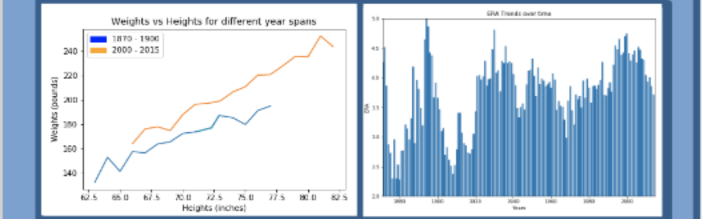
Introduction

Since the late 1800s, Major League Baseball (MLB) has seen multiple new trends emerge. For pitchers, in the last fifteen years, ERA's (earned run average) have decreased more than 20%; meanwhile, hitters' batting percentages and home runs have decreased by about 7% and 24%, respectively. Dominant players obviously have characteristics that distinguish themselves from others in the MLB, but are those characteristics physical ones or related to a player's individual skill set? In this project, we examine what those characteristics are. Specifically, we will focus on examining physical attributes of MLB players as well as skill-based characteristics, and determining which ones are more related to elite, Hall of Fame players.

Objectives

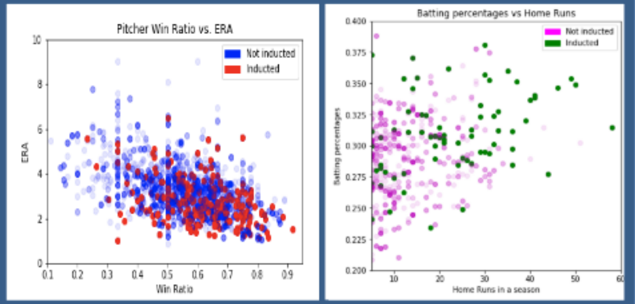
1. Analyze historical progression in physical/skill based attributes to predict future progression
2. Label data related to physical attributes and skill set; designate which characteristics were more conducive to having a Hall of Fame career
3. Predict and execute optimal classification algorithm for dataset

Overarching Trends



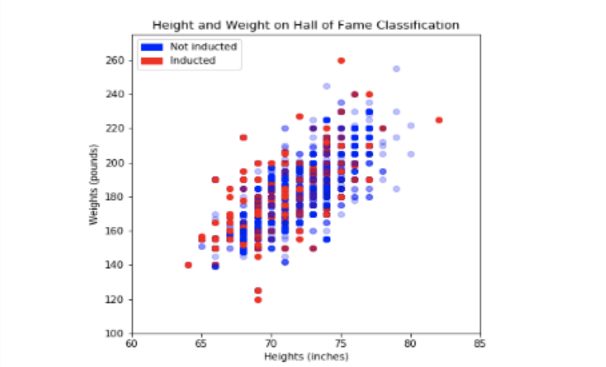
In the last 15 years, pitcher ERA's have decreased significantly, an indication of an increasing skill level of players. At the same time, heights and weights of players have also shown significant jumps since the advent of baseball.

Skillset Labeling



Plot of win ratio vs ERA (for pitchers) shows that higher win ratios and lower ERA's make a player more likely to be inducted. A plot of home runs per season (HRS) vs Batting percentages shows that players with higher HRS's and batting percentages are more likely to be inducted, as we expect.

Physical Attribute Labeling



Plot of height vs weight shows that in general, from a physical standpoint, there was no clear distinction between physical attributes for Hall of Fame and non Hall of Fame players.

Logistic Regression

The table below shows the coefficients of all variables after performing logistic regression. The following observations are made:

1. Coefficients for physical attributes (height, weight) range between 0.003 - 0.04. Based on this, these values have a minimal impact on our probability function
2. Coefficients for skill based characteristics (ERA, Batting %, etc) range between 0.1 - 0.6. These values are higher and have a larger impact on our probability function.

	Pitchers	Hitters
Heights	0.043	0.065
Weights	0.003	0.009
ERA	0.109	-
Win Ratio	0.360	-
Batting %	-	0.667
Home runs/season	-	0.123

Classification Algorithms

Objective: We see above that skill sets are stronger determining factors in Hall of Fame designation than physical attributes. But, to what degree is this true? Below we use various classification techniques to determine this.

1. Consider all variables height, weight, ERA and win ratio (for pitchers) and height, weight, batting % and home runs per season (for batters)

2. Try multiple classifiers two scenarios: inducted and not inducted into the Hall Of Fame

3. Score each classifier between k-neighbors, LDA, Logistic Regression, Decision Tree, Random Forest, Naive Bayes and QDA. Choose the best method based on scores.

Pitching Data

	Score of each classifier	
Classifier	K Neighbors	0.89
	LDA	0.92
	Logistic Regression	0.94
	Decision Tree	0.90
	Random Forest	0.93
	Naïve Bayes	0.92
	QDA	0.91

Batting Data

	Score of each classifier	
Classifier	K Neighbors	0.92
	LDA	0.89
	Logistic Regression	0.95
	Decision Tree	0.93
	Random Forest	0.92
	Naïve Bayes	0.91
	QDA	0.90

Conclusions

1. As expected, compared to physical attributes, skill characteristics seem to be very highly associated with Hall of Fame induction
2. These data suggest that in the future, although physical attributes may play a small role in a player's success, players should focus more on improving skill-based characteristics such as hand-eye coordination and stamina
3. Logistic Regression suggests win ratio and batting percentage to be most significant of variables analyzed, and was consistently the best classifier for our set

References

1. King, Ed, "Baseball Databank", *Kaggle*, www.kaggle.com/open-source-sports/baseball-databank, Web.
2. "Timeline of Baseball", *PBS*, http://www.pbs.org/kenburns/baseball/timeline/, Web.