Who's on Deck for Cooperstown?

James Tang & Guy Dotan

Agenda

1. Introduction
2. Our Dataset
3. Exploratory Data Analysis
4. Our Models
5. Conclusions/Predictions
6. Further Research

Introduction

"Club-329"

- Have been 19,000+ players in the 150-year history of pro baseball
- Only 261 players elected to the Hall of Fame (~1%)
- Additional 68 people with other roles

Member Breakdown

- BATTERS - 178
- PITCHERS - 83
- MANAGERS - 23
- UMPIRES - 10
- EXECUTIVES - 35

Balloting Process

- 5 years after retirement, player is eligible
- Members of BBWA (~300-500) vote up to 10 players
- Must appear on 75% of ballots to be inducted

Our Dataset

[Lahman Baseball Database]

- Freely available MLB dataset with records dating back to 1871
- We only used **PITCHERS**
  - Defined by 1000+ regular season innings pitched in career
- Removed players whose career ended before 1900
- 71 total variables…
  - 5 player descriptors → (name, first year, last year, etc.)
  - 25 regular season stat → (innings, losses, strikeouts, etc.)
  - 24 postseason stats → (same as before but in playoffs)
  - 5 awards → (All-Star, Cy Young, MVP, etc.)
  - 1 Sabermetric stat → (WAR - Wins Above Replacement)
  - 10 season-rank stats → (Top 10 ERA, Top 5 Wins, etc.)
  - Inducted in HOF → (Response: Yes or No)

Feature Motivation

- Starting Pitchers vs Relief Pitchers
  - Starters tend to pitch more innings and therefore have larger total counting stats e.g. Strikeouts
  - Starters: Wins
  - Relievers: Saves
- Career Longevity vs Career Excellence vs Single Season Excellence
  - Career Longevity: Total Wins, Total Saves
  - Career Excellence: Career ERA
  - Single Season: Top 5 in Wins, All star seasons, Cy Young awards
- Advanced Stats vs Traditional Stats
  - Advanced: WAR (Wins Above Replacement)
  - Traditional: Wins, Saves, ERA, Strikeouts

WAR! What is it Good For? … actually a lot

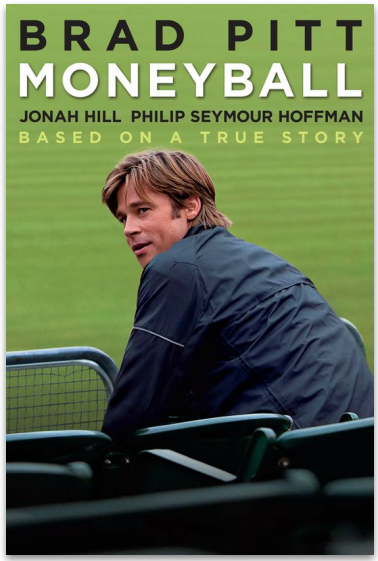
- What's problematic with traditional stats?
  - Other factors outside of player performance affect traditional stats.
  - Wins are influenced by your team's and opponent's offense.
  - ERA is influenced by team defense, opponent's offense, park, DH, etc.

WAR (Wins Above Replacement)

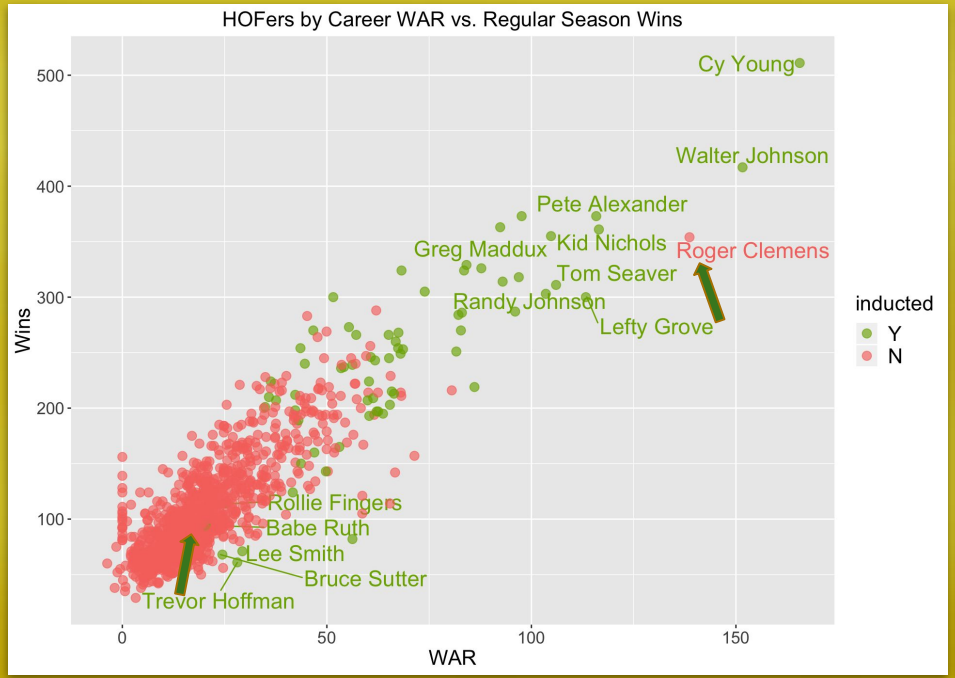
- What's problematic with traditional stats?
- Provides a single metric of player value
- Context, league, and park neutral
- Can use WAR to compare players between years, leagues, and teams

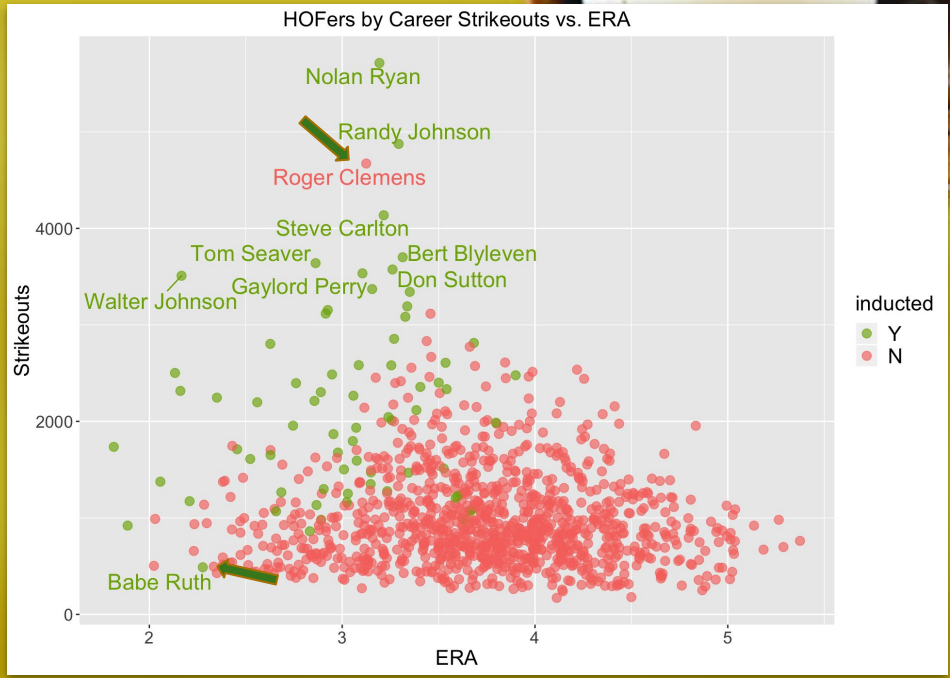
Statistical analysis to transform:

Traditional stats → Wins contribution

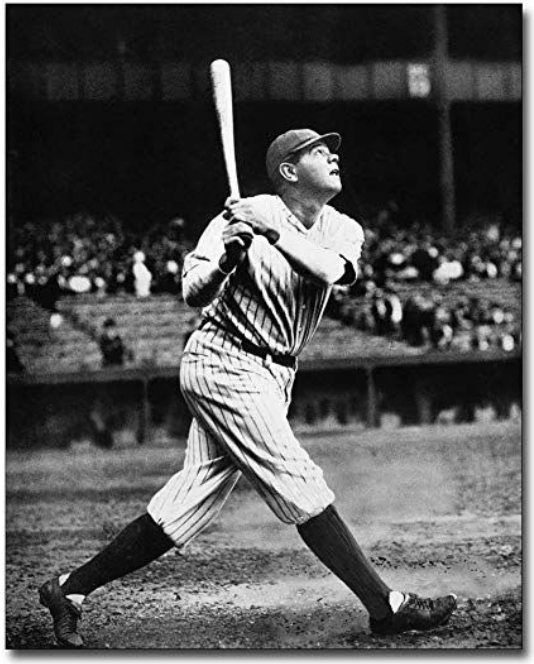


Exploratory Data Analysis





Removing Outliers

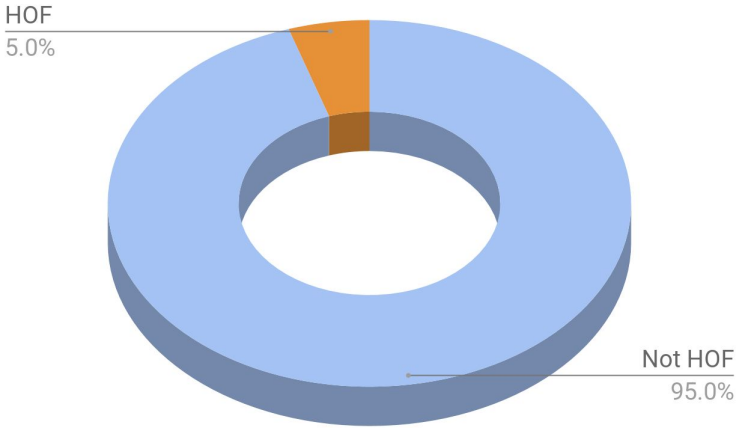


Babe Ruth - primarily a batter

Roger Clemens - steroids scandal

Training and Test data

- Imbalanced data (6.6% HoF)
- **Oversampled** Hall of Famers
- Training data (29.8% HoF)
  - 95% non-HoF data
  - 70% HoF data replicated 8x
- Test data (29.8% HoF)
  - 5% non-HoF data
  - 30% HoF data



Model 1

Structure

- No interactions
- GLM binomial

Selection

- step() function
- BIC

Model 1 - Coefficients

Variable	Coefficient
total wins	0.111
total losses	-0.116
total saves	0.065
career ERA	-8.249
total ER	0.014
total strikeouts	-0.003
mvp awards	2.697
top10 saves	-0.362
top10 strikeouts	0.743

Observations

- In general, positive coefficients for good outcomes and negative coefficients for bad outcomes
- Non-intuitive coefficients
  - Negative for Top10 in Saves
  - Negative for Total Strikeouts
  - Positive for ER
- Career Longevity vs Single Year Excellence
- WAR ignored. WAR is a recent invention. Model detects that HOF voters used traditional stats as HOF criteria.

Model 1 - Diagnosis

TRAINING DATA

		Actual: No	Actual: Yes
Predicted	No	878	8
	Yes	24	376

Accuracy: 97.5%

TEST DATA

		Actual: No	Actual: Yes
Predicted	No	47	2
	Yes	0	18

Accuracy: 97.0%

Model 2

Structure

- Include square features and interactions
- GLM binomial

Selection

1. Among all possible features, square features, and interaction features, pick the one which lowers deviance the most.
2. Repeat until we cannot lower deviance by at least 25

Judgement call: we try to pick a model with intuitively correct coefficients and good out-of-sample accuracy and good deviance.

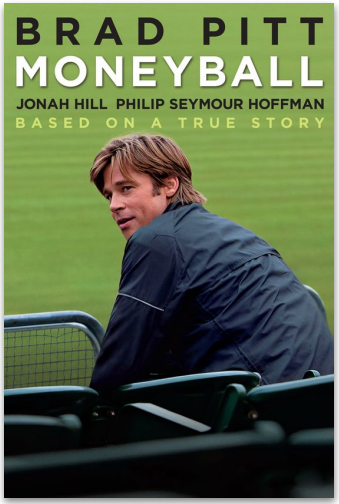
Model 2 - Coefficients

Variable	Coefficient
total_WAR:career_ERA	0.010
all_star_seasons:total_saves	0.006
top5_wins:career_ERA	0.093
total_wins:career_ERA	0.013
total_losses:career_ERA	-0.008

- Variables are ordered from most important (top) to least important (bottom) in terms of lowering deviance
- We will see how the model successively addresses different criteria for Hall of Fame

Model 2 - Most Important Feature

Interaction: `career_WAR` and `career_ERA`



- Advanced Stat (WAR)
  - With access to only one feature, the model picks our "single metric of player value"
- Rewards Career Longevity (Total stat) and Career Excellence (ERA)
- Favors Starting pitchers (Total stat)

Model 2 - 2nd Most Important Feature

Interaction: `Total_Saves` and `All_Star_Seasons`

- Favors Relief Pitcher (saves)
  - Our model addresses the blindspot of the first feature
- Rewards Career longevity (total stat) and Single season excellence (All-Star seasons)

Model 2 - 3rd Most Important Feature

Interaction: `Top5_wins` and `career_ERA`

- Traditional Stat (Wins)
  - Model complements the first feature with traditional stats, which is what HOF voters probably judged by
- Favors Starting Pitchers (Wins)
- Rewards single-season excellence (Top 5 in wins) and career excellence (ERA)

Model 2 - Diagnosis

TRAINING DATA

		Actual: No	Actual: Yes
Predicted	No	882	8
	Yes	29	376

Accuracy: 97.8%

TEST DATA

		Actual: No	Actual: Yes
Predicted	No	47	2
	Yes	0	18

Accuracy: 97.0%

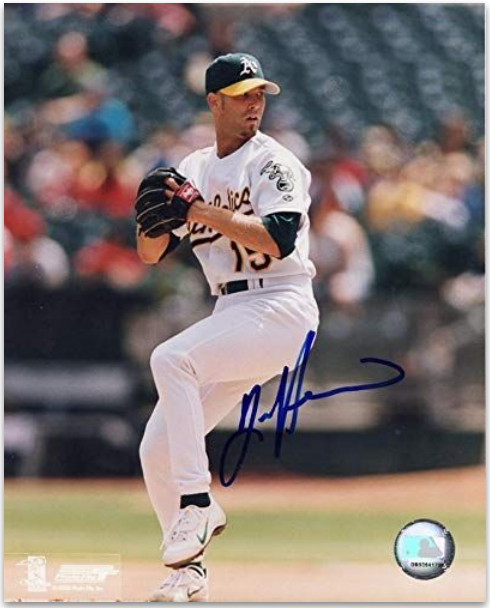
Model Comparison

- Same accuracy on test data (same confusion matrix)
- Model 2: slightly higher deviance but fewer features
  - Model 1: deviance 173.4 on 1276 degrees of freedom
    - Net_deviance / net_df = 273.60
  - Model 2: deviance 200.05 on 1280 degrees of freedom
    - Net_deviance / net_df = 154.96
  - Null Model: deviance 1568.08 on 1285 degrees of freedom
- Model 2 reveals feature importance rankings

Prediction on Upcoming Nominees

(Players who retired in 2015 or later)

Active Pitcher	HOF Odds	
Tim Hudson	100.0%	✔
Justin Verlander	46.2%	?
Clayton Kershaw	32.8%	✘
Mark Buehrle	26.6%	✘
Bartolo Colon	26.5%	✘
Zack Greinke	7.6%	✘
Adam Wainwright	5.3%	✘
Felix Hernandez	3.7%	✘
Jared Weaver	2.6%	✘
Max Scherzer	1.0%	✘

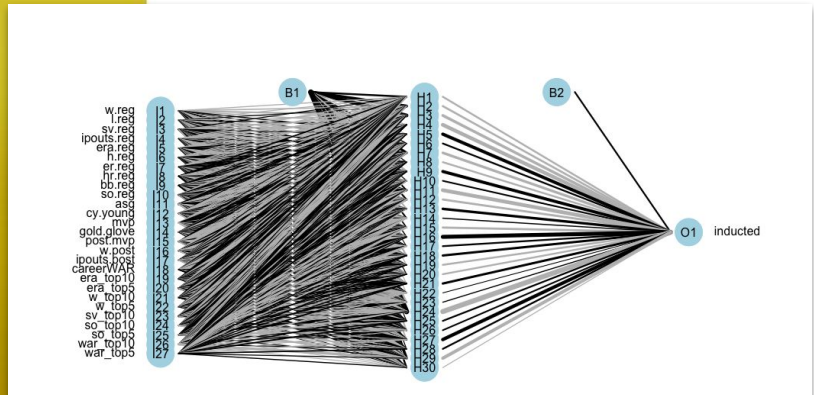


Bonus Model - Neural Network

- Normalized data | 70:30 train:test | 1 hidden layer w/ 30 nodes

		Actual: No	Actual: Yes
Predicted	No	285	3
	Yes	7	17

Train/Test Accuracy 96.8%



Active Pitcher	HOF Odds	
Bartolo Colon	100.0%	✓
Tim Hudson	100.0%	✓
Justin Verlander	100.0%	✓
C. J. Wilson	100.0%	✓
Felix Hernandez	99.7%	✓
Max Scherzer	96.8%	✓
Clayton Kershaw	77.0%	✓
CC Sabathia	63.7%	✓
Adam Wainwright	5.1%	✗
James Shields	2.0%	✗

Further Research

Hitters

More Sabermetrics

Compare Different Eras

Ballot Votes

THANK YOU!