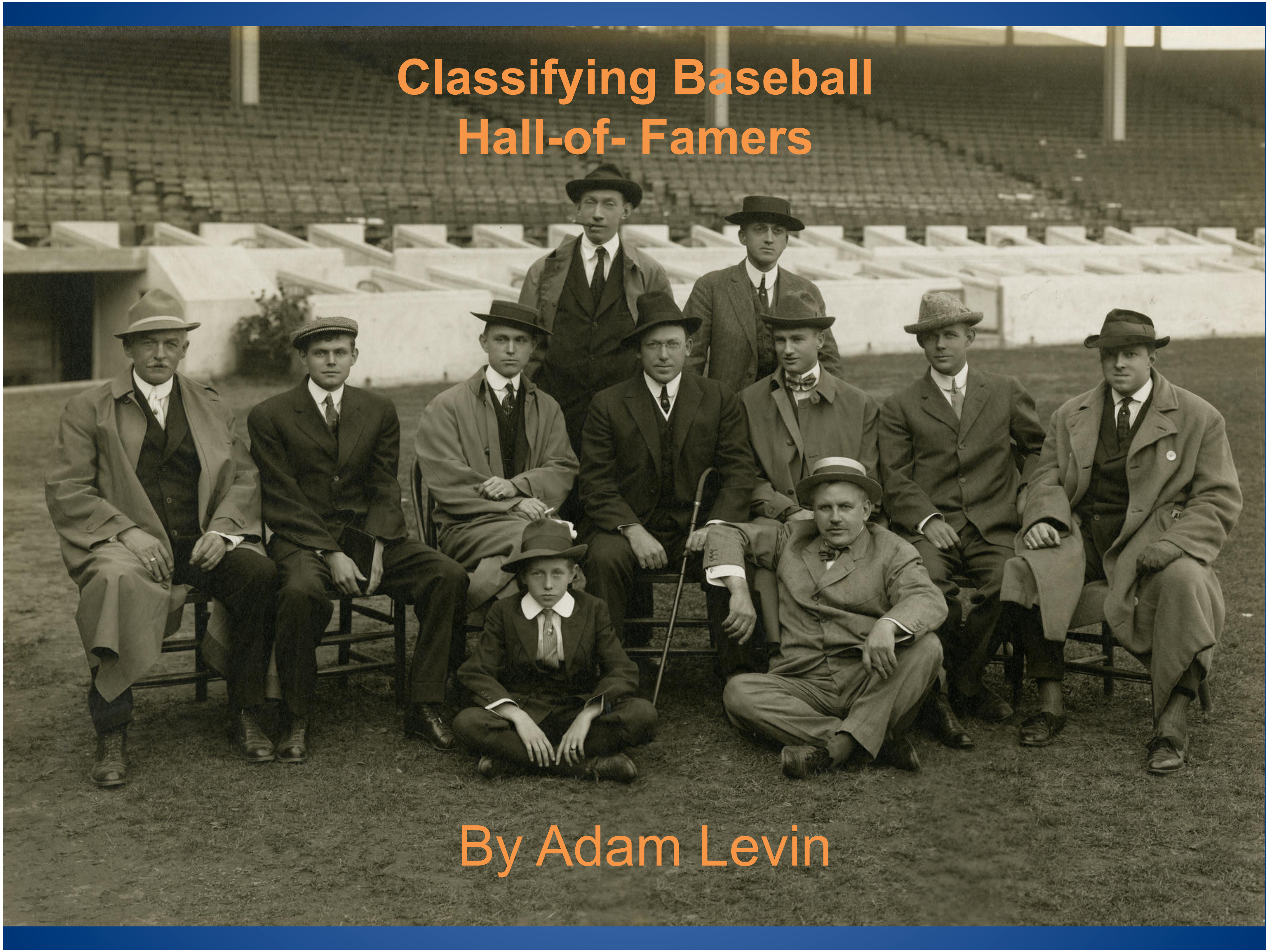
Classifying Baseball Hall-of- Famers



By Adam Levin

The Problem

- We see the first 10 years of an MLB player’s career, we try to determine whether or not he’ll make it to the hall of fame
- Data
    - Principally Sean Lahman’s Baseball Database
    - Limit to players who have been eligible for the hall of fame for at least 5 years

Actually, two problems

- Positional players (everyone but pitchers)
  - 146 HOF players out of 8363 (1.7%)
- Pitchers
  - 62 HOF players out of 6900 (0.9%)

Feature Engineering

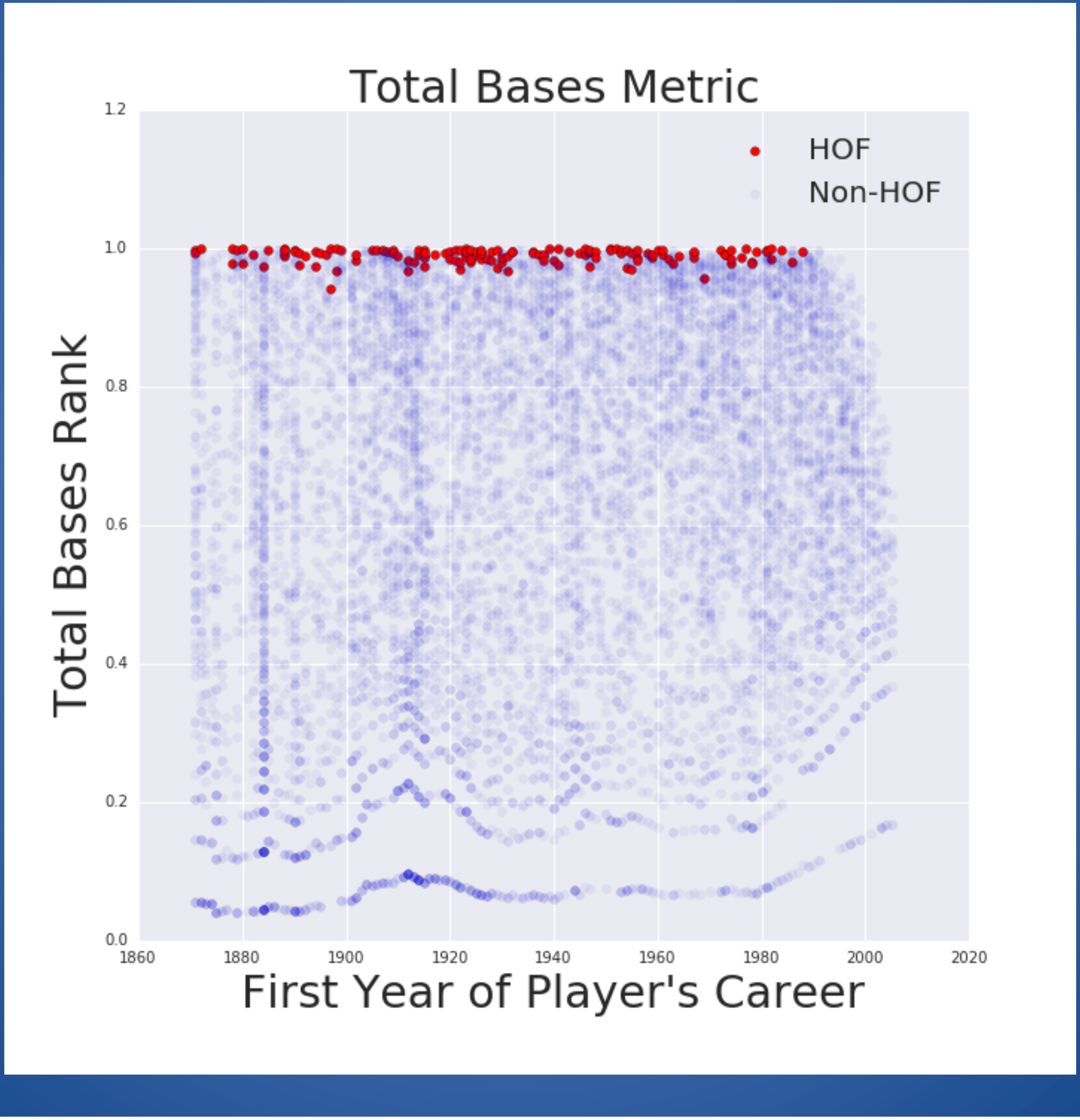
- The basics like hits, at bats, homeruns for hitters; wins, strikeouts, shutouts for pitchers

- Robust proportion features:
  - Home runs per at bat for batters
  - Strikeouts per batter for pitchers

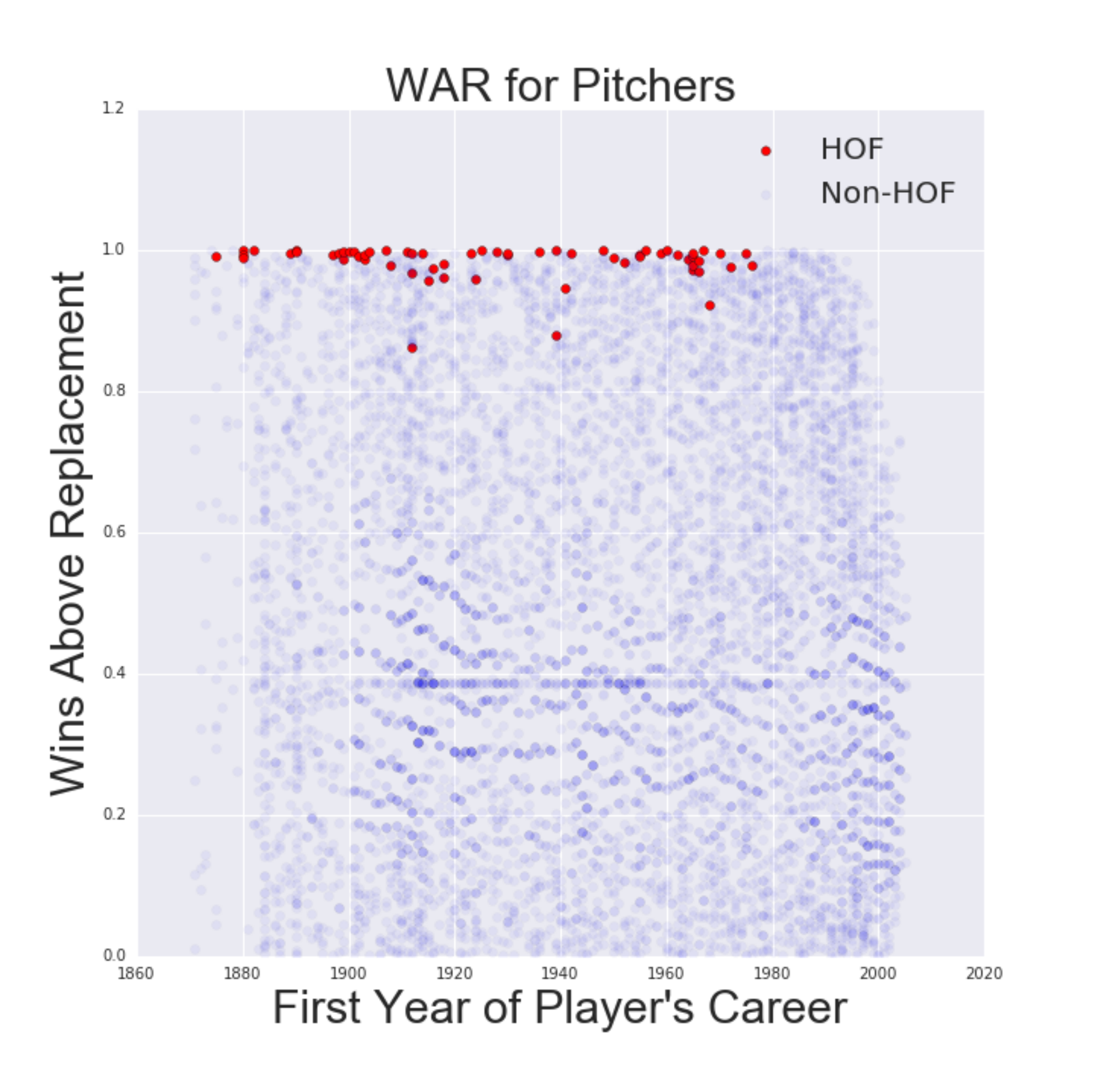
- Awards like MVP, Cy Young along with voting on awards

- Wins-Above-Replacement

An example batting statistic



An example pitching statistic



Modeling

- Criteria Optimized for:
  - Area under the precision-recall curve

- Grid Search to tune parameters

- Best models:
  - SVM with rbf kernel
  - Gradient Boosted Classifier

- Voting classifier to ensemble the two

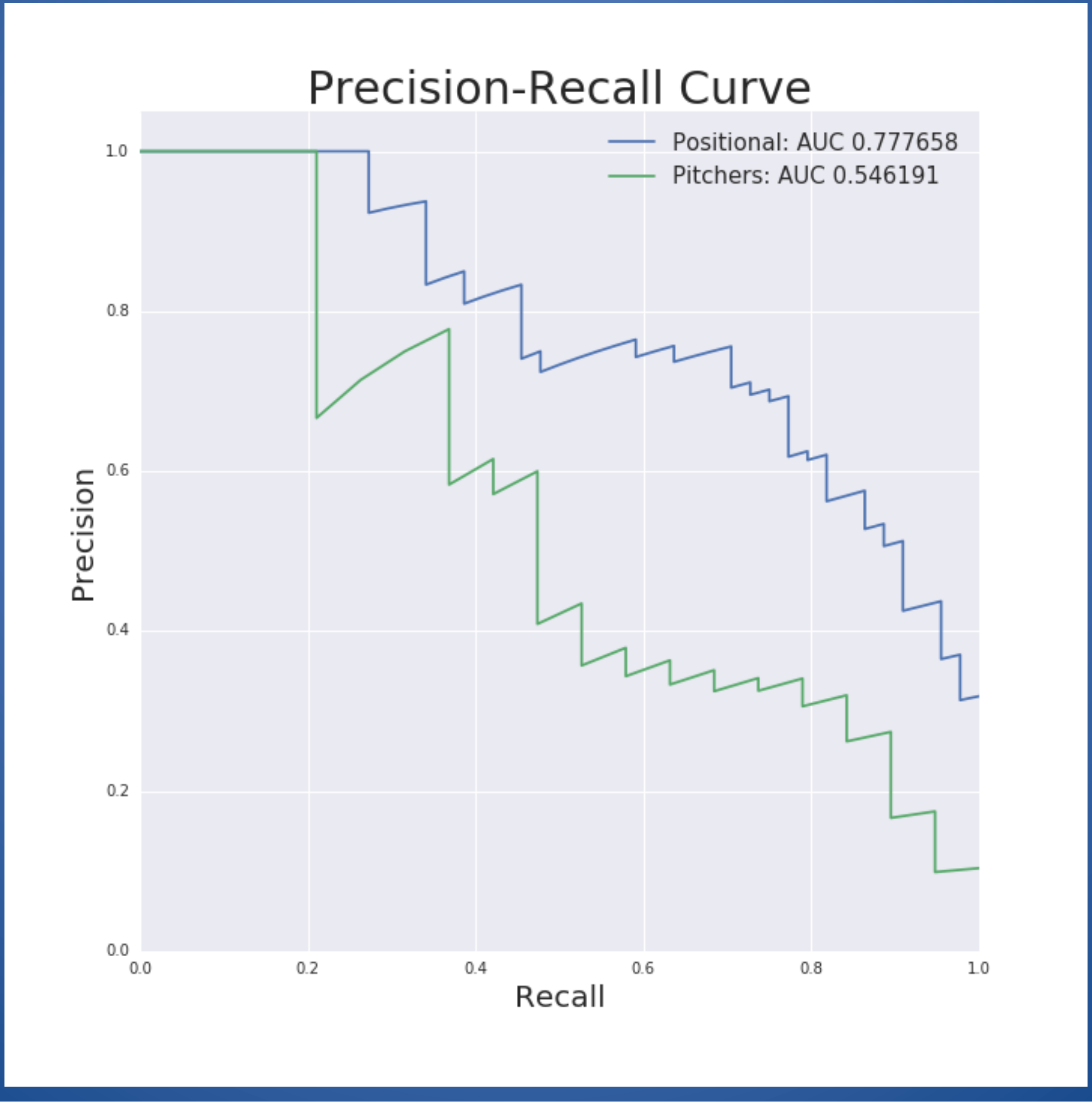
Results

	Training AUC	Testing AUC
Positional	.747	.738
Pitchers	.622	.604

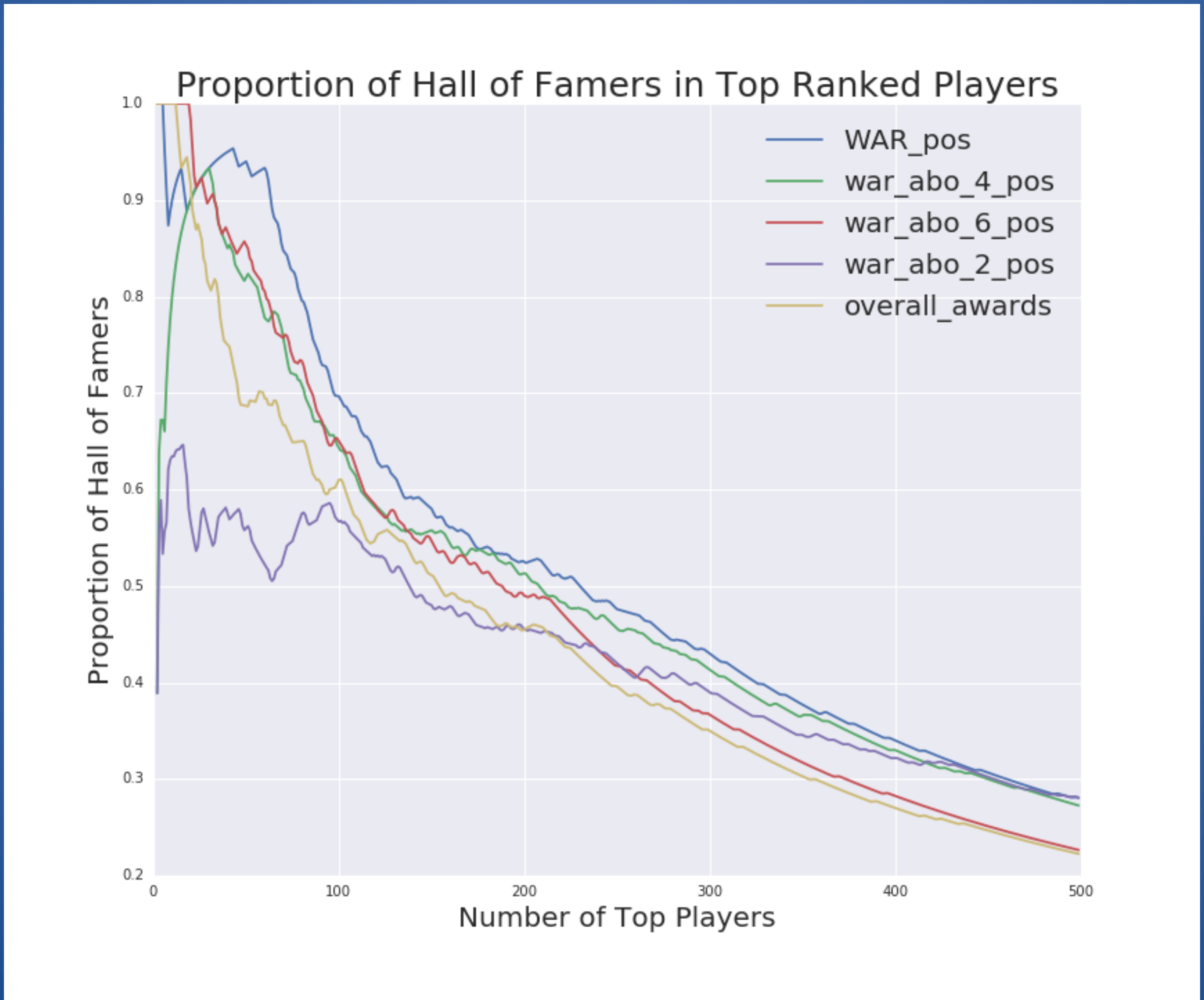
Positional	HOF	Non-HOF
Predicted HOF	40	39
Predicted Non-HOF	4	2426

Pitchers	HOF	Non-HOF
Predicted HOF	16	39
Predicted Non-HOF	3	2012

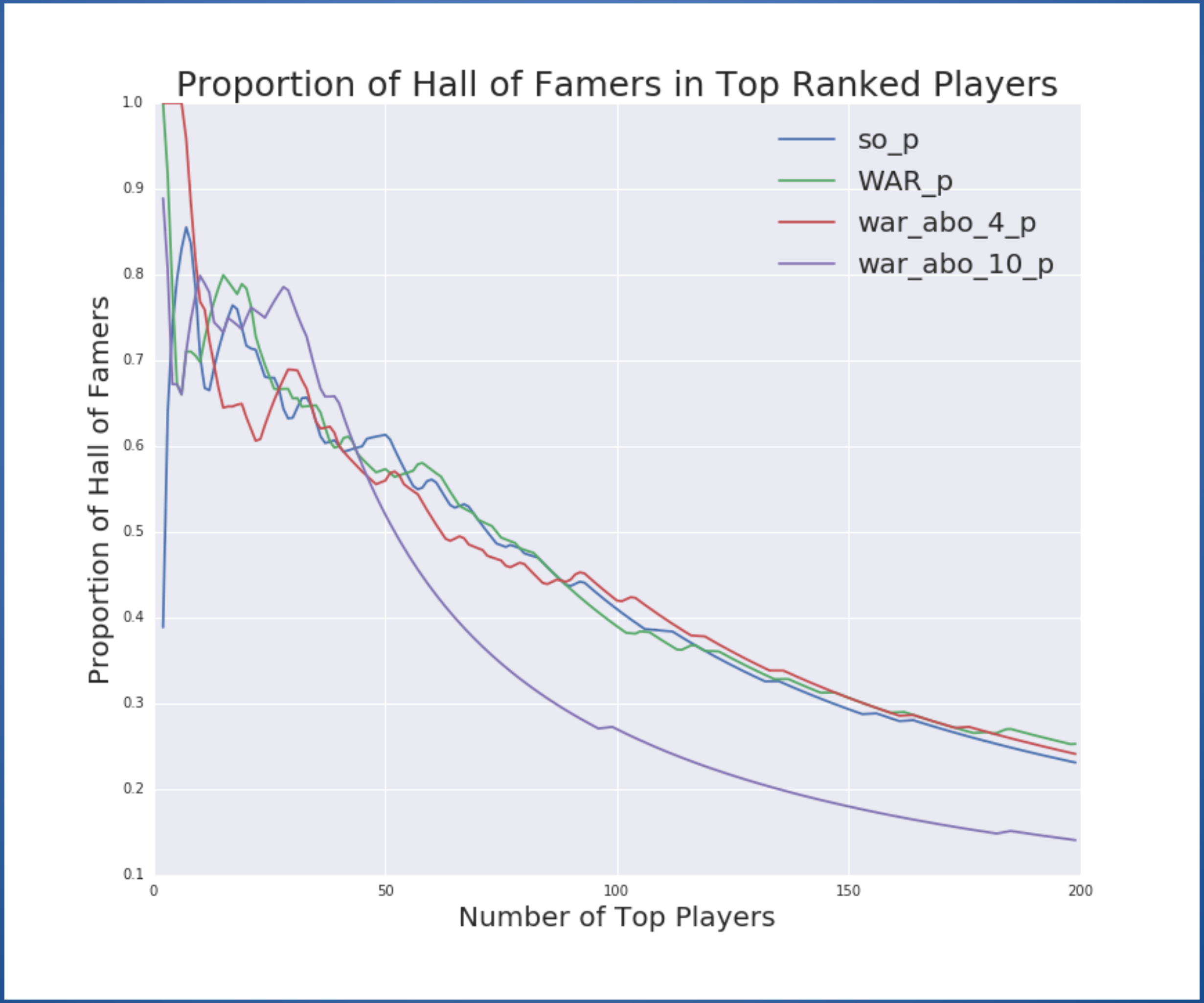
Recall-Precision Curve



Most predictive features for Positional Players



Most predictive features for Pitchers



Thank you