PREDICTING BASEBALL LEGENDS

ELIZABETH HARWOOD

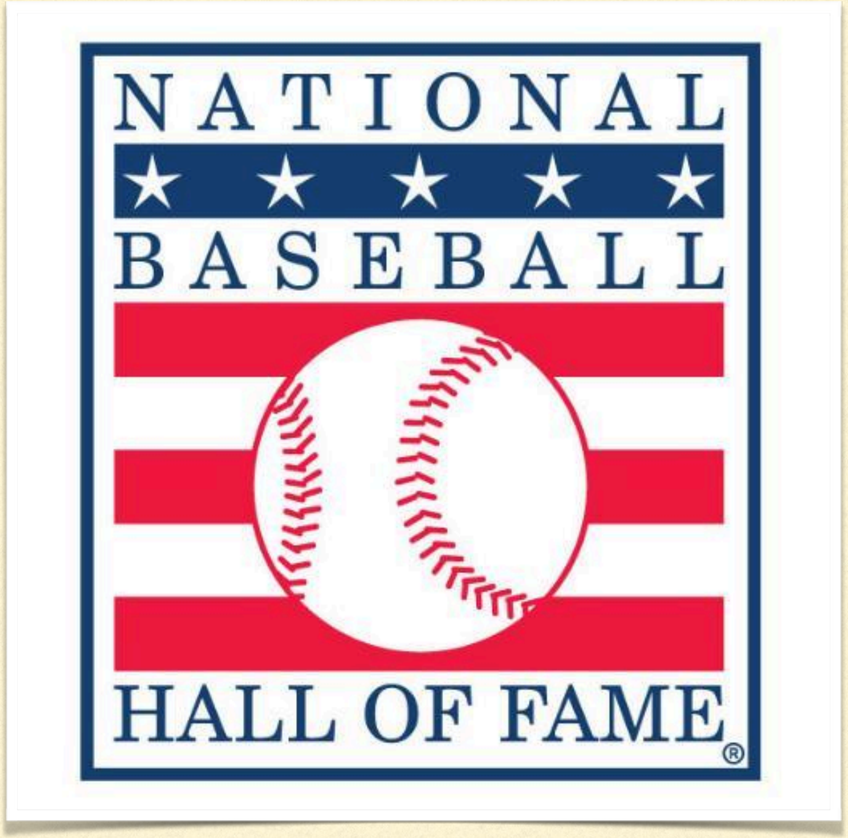
BACKGROUND - BASEBALL HALL OF FAME

Then-current members in 1939 (http://baseballhall.org/)

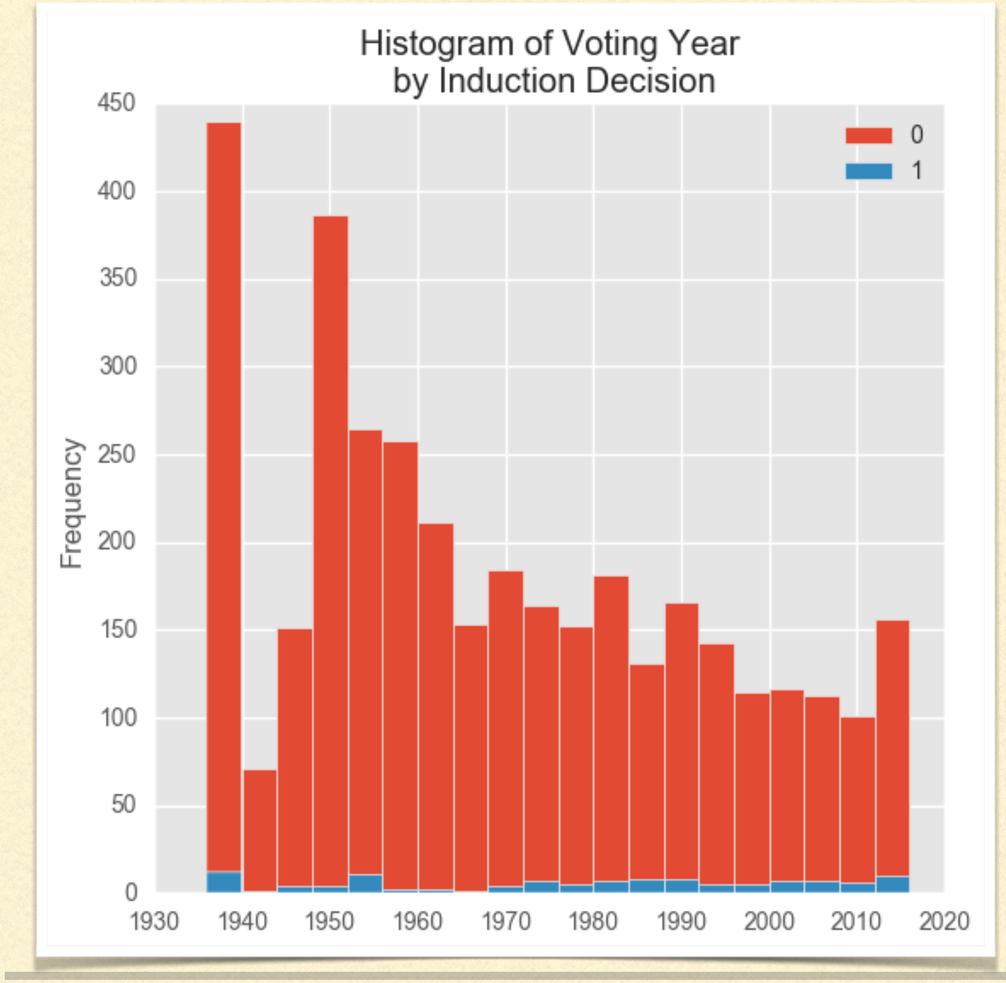
- Baseball Hall of Fame - Since 1936, 317 members currently
- Inductions by category:
  - 220 ML players, 30 executives, 35 Negro Leaguers, 22 managers, 10 umpires
- Individuals must receive vote on 75% of ballots cast in order to be chosen for induction

PROBLEM SUMMARY - PREDICTING HALL OF FAME INDUCTION

- Predicting whether or not a nominee for induction to the Baseball Hall of Fame will be chosen for induction
- Which factors most influence induction into the hall of fame?
- Which factors are less predictive?
- Using data from the Baseball Databank, curated and maintained by Sean Lahman
  - http://www.seanlahman.com/baseball-archive/statistics/



DATA AVAILABLE



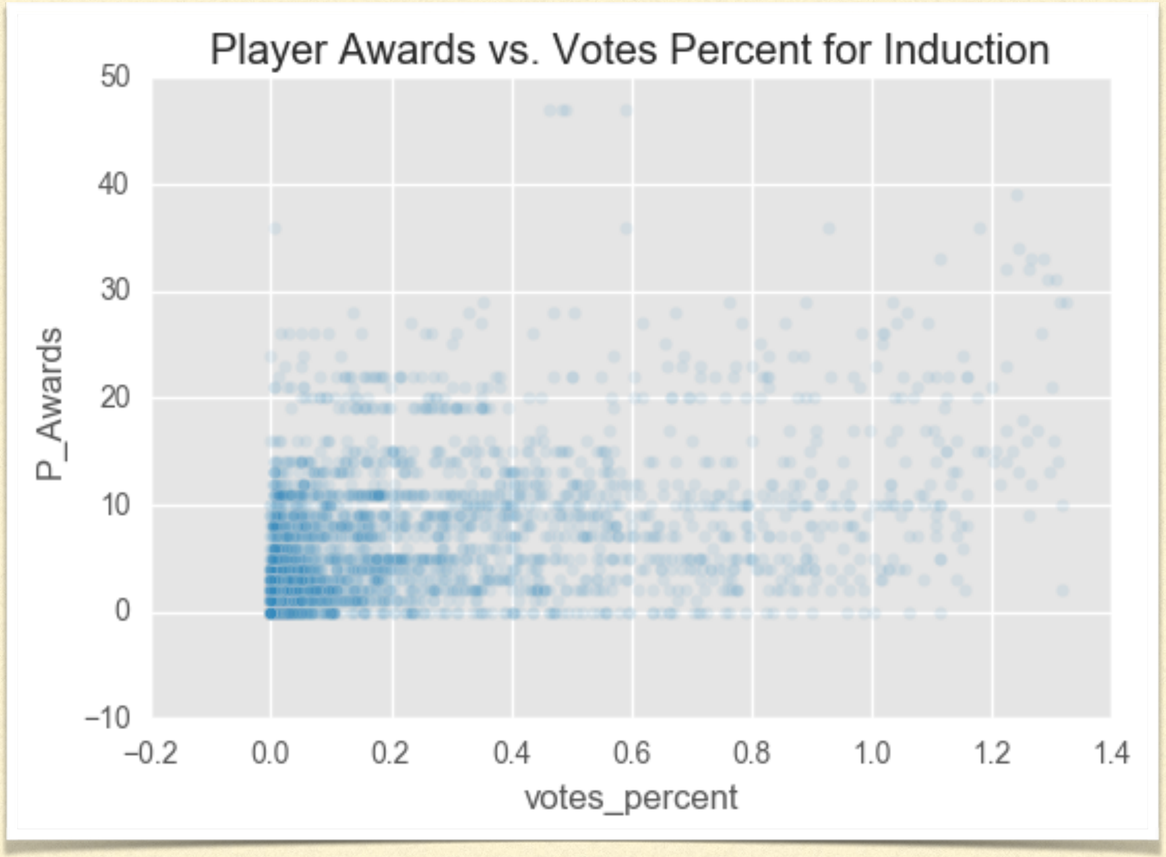
- Primary data table: Hall of Fame Induction decision for each nominee (along with vote breakdown for majority of votes)
- Additional Data in the dataset includes:
  - Players fielding
  - Team Record by season
  - Individual player statistics (batting and fielding)
  - Salary by player and year
  - Individual Awards given

DATA INSIGHTS

- Every data point has binary ‘inducted’ for outcome
- Most points have breakdown of votes, allowing for creation of a second, continuous outcome, ‘votes_pct’ (% of needed votes received)
- Two sets of categorical data have a large impact on rate of induction success: ‘votedBy’ (type of vote), and ‘category’ (player, manager, umpire, etc)
- Additional individual and team statistics must be aggregated and included for each nominee

INDUCTED	0	1	SUCCESS RATE
votedBy			
BBWAA	3573	116	0.031445
Centennial	0	6	1.000000
Final Ballot	21	0	0.000000
Negro League	0	26	1.000000
Nominating Vote	76	0	0.000000
Old Timers	0	30	1.000000
Run Off	78	3	0.037037
Special Election	0	2	1.000000
Veterans	60	129	0.682540

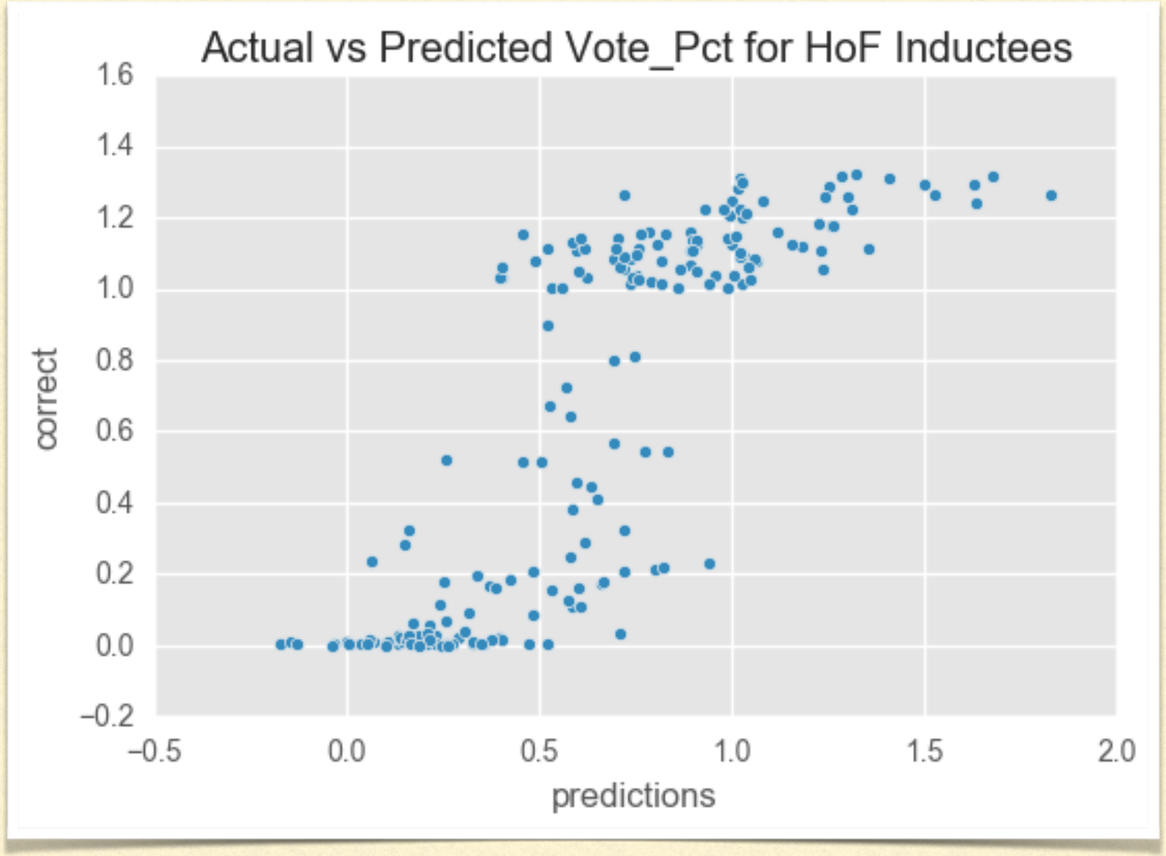
FEATURE ENGINEERING



- Features engineered:
  - Votes_pct (outcome variable, % of votes needed for induction that an individual received)
  - Salary (max, min, sum) over career
  - Number of All-Star games participated in (since 1933)
  - Batting Average, Number of seasons in top 10% batting average, Total Runs over career
  - Number of Strike Outs, Shut outs and opponents' batting avg (for Pitchers)
  - Pitcher (binary if player is pitcher or not)
  - One-hot variables for Category (Player, Ump, Manager, etc)

MODELING APPROACH

- Logistic regression with binary Inducted 0/1
- Linear regression with Votes_Pct variable
- Feature selection with Lasso Regression
- Decision Trees with feature importance ranking
- Test set scoring to choose final model, re-fitted with entire data set



CHALLENGES

	0	1	RATE
INDUCTED	3497	116	0.0321

Data with Null Values:

Wins, batting statistics,
pitching statistics,
salary statistics

- Some features had many null values due to availability of data
  - Ex: Salary dropped due to null values for 83% of points, data only available since 1985
- Deciding which data to include - Risks include removing data that has some values missing and removing a disproportionate amount of one outcome or the other, thus reducing the model's ability to predict the outcome
- Disproportionate outcome of '0' for induction required downsampling for binary classification

DEALING WITH NULL VALUES

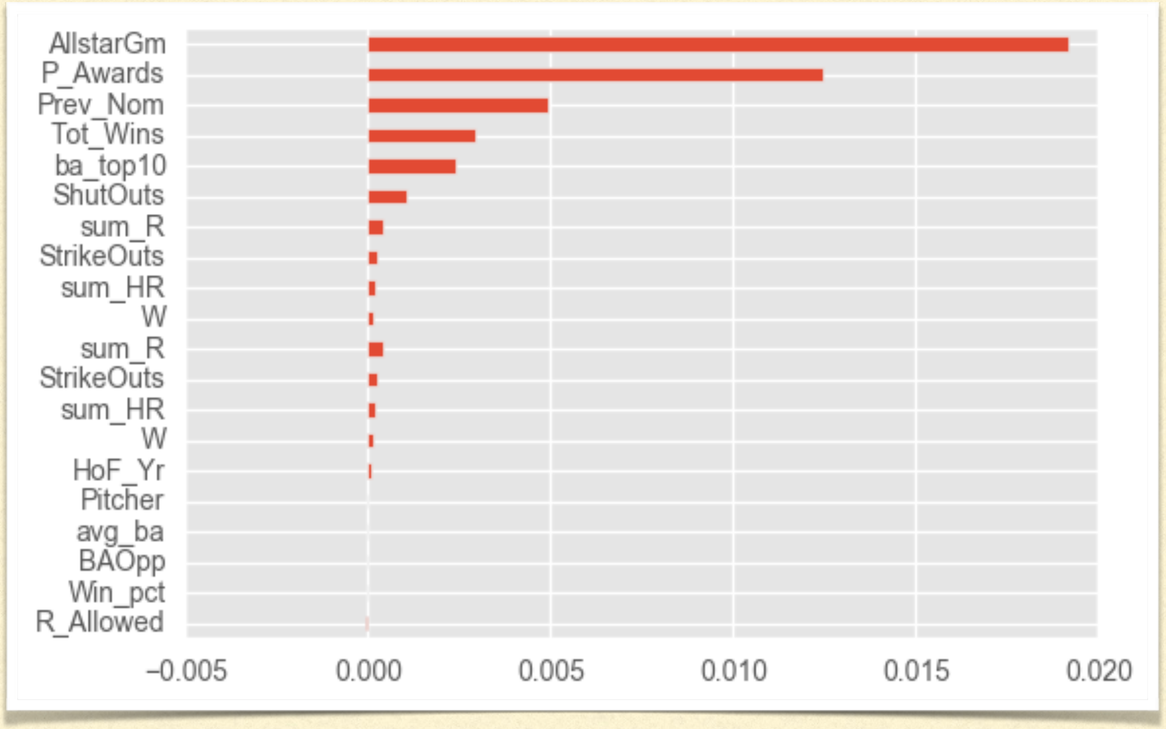
- Features for one category were often null for other category individuals. Examples:
  - Set Pitcher to 0 if the inductee is anything but a player
  - Pitching statistics set to 0 if null and Pitcher == 0, same for converse with batting statistics if individual is a pitcher or not a player
- All star appearances null for half of nominees, we have data since 1933 (is it complete?)
  - Missing data is mostly for players, even after 1960; assume those players were not in any all star games and set missing vals to 0

DOWN-SAMPLING:

- Took an equal number of successful and unsuccessful induction nominees for model training
- Greatly improved accuracy of predictions
  - more apparent in logistic regression models

SELECTED RESULTS

Feature Importance with Lasso Regression



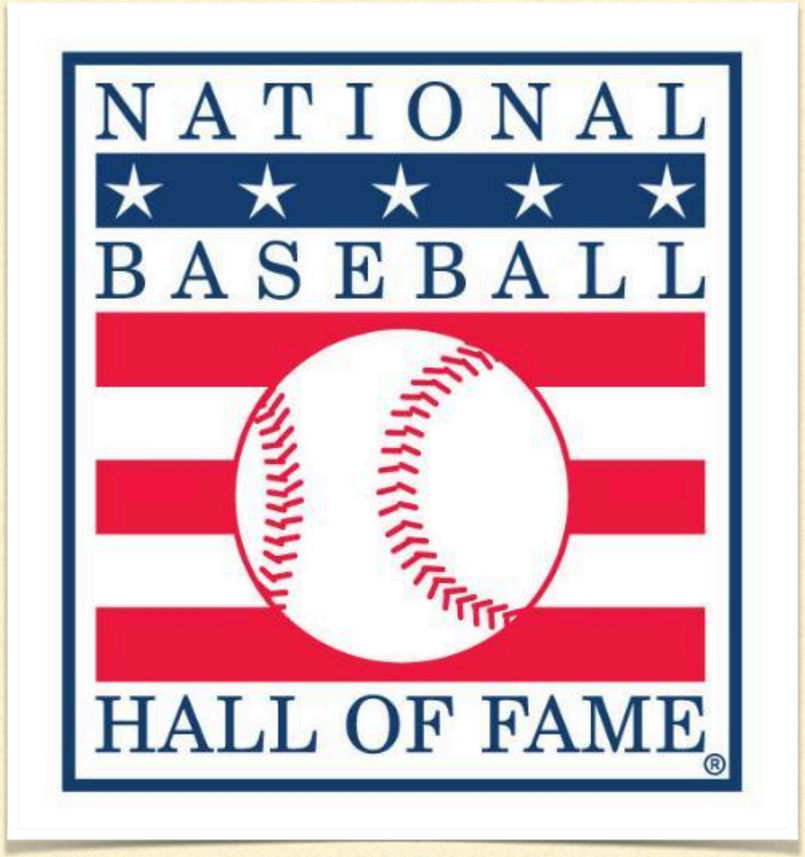
- Most important features:
  - AllstarGm: number of all star games chosen for
  - P_Awards: number of individual awards received
  - Prev_Nom: number of previous years as a nominee for Hall of Fame
- Train RMSE: 0.320639595285
- Test RMSE: 0.403389820585

FINAL MODEL

CONFUSION MATRIX FOR TRAINING DATA	TRUE INDUCTED	0	1
HYPOTHESIZED INDUCTED			
0		92	19
1		24	97

- Logistic Regression with features:
  - 'P_Awards': number of individual awards given to a player
  - 'StrikeOuts': strike outs by a pitcher
  - 'ba_top10': number of seasons with a top 10% batting average
  - 'sum_R': sum of career runs
- Test accuracy rate: 0.833 (on unseen data), created with all available data

CONCLUSIONS



- Trying to optimize for accuracy (or prediction error) in logistic regression was a more understandable outcome
- Vote_pct was a proxy outcome variable, so it didn't fit the problem statement as well
- Individual Statistics were more predictive of final outcome than team statistics
- Difficult to stay organized during iterative data cleaning/ feature engineering and modeling

NEXT STEPS

- Separate Models by Inductee Category, Election Type (‘voted_by’)
- Decide whether to optimize for Precision or Recall
- Predict next batch’s nominee chances

QUESTIONS?

APPENDIX: MODELING RESULTS

- Logistic reg (chosen for final model): 0.8326 = test data accuracy
- Final model:
- ['P_Awards', 'StrikeOuts', 'ba_top10', 'sum_R']
- [0.13839703 0.00123415 -0.08594097 0.00210337]
- Decision Trees Results: (0.592827729967637, 'P_Awards'),
  (0.14652941101724637, 'StrikeOuts'),
  (0.06817645671488517, 'ba_top10'),
  (0.04673159662543950 6, 'sum_R'),
  (0.04140127388535031 6, 'Tot_Wins'),
  0.722 Accuracy score on Test data

MODELING RESULTS:

BASELINE MODEL

- Linear Regression with Statsmodels
- All Features Included
- Adjusted R-squared: 0.480, F-statistic: 100.8

MODEL WITH DOWNSAMPLING

- Linear regression and statsmodels:
- R-Squared: 0.757, F-statistic: 239.4
- Statistically Significant Coefficients: AllstarGm: 0.0415, Prev_Nom: 0.0198, P_Awards: 0.0256