Getting to Cooperstown

Predicting the Baseball Hall of Fame Vote

A Successful Startup Needs A Creation Myth

Baseball Hall of Fame founded in Cooperstown, NY about 100 years after it was not invented there by a future Civil War officer

Baseball writers vote based on: *“player's record, playing ability, integrity, sportsmanship, character, and contributions to the team(s) on which the player played.”*

Need 75% of the vote to be inducted and can appear for up to 10 years on the ballot. Can vote for up to 10 players per ballot

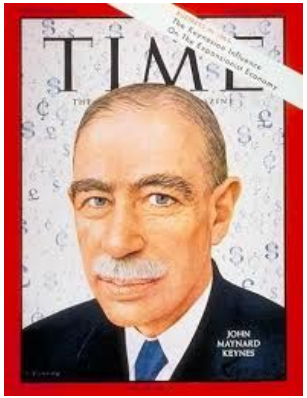
Two types of people in this world.....

Why would anyone care about the Hall of Fame?

Baseball people

Long lineage of baseball in American popular culture

Nature of game makes it (fairly) easy to separate individual performance. .. Fuel for arguments!



Metis people

Predicting a heuristic system like a physical or economic one

"*The ideas of economists and political philosophers, both when they are right and when they are wrong are more powerful than is commonly understood. Indeed, the world is ruled by little else. Practical men, who believe themselves to be quite exempt from any intellectual influences, are usually slaves of some defunct economist.*" - - *Keynes*

Dead pull hitter to right field

Classification - > Prediction by targeting percentage of vote received on first year on ballot

Focused on batters, no pitchers

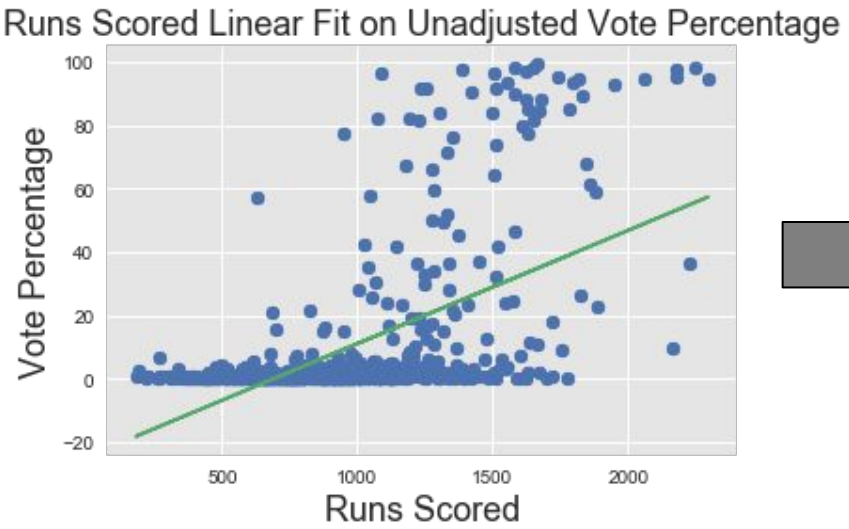
Highly skewed target (665 players, only 131 had more than 5% of the vote)

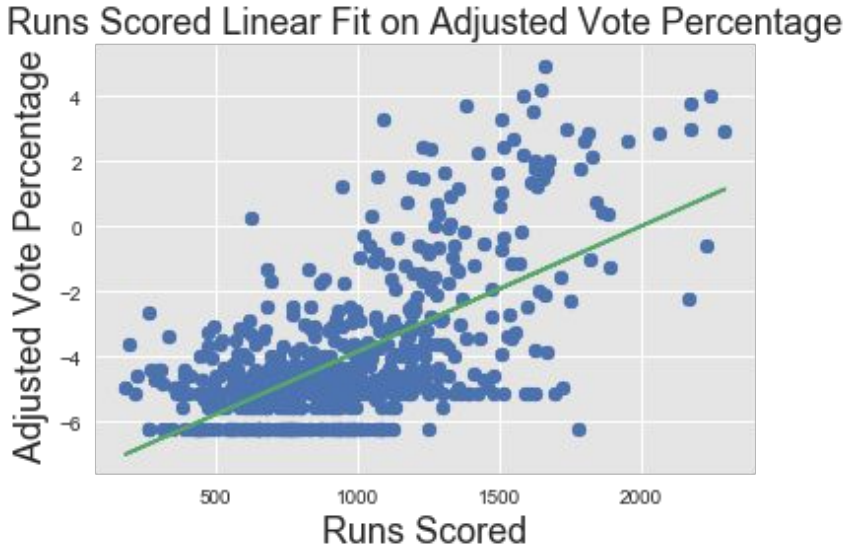
Know the target is capped and floored, can’t go below 0% of the vote and can’t go above 100% of the vote

Target Transformation

Predicting vote % received transformed to predicting log of [vote% / (100 - vote%)]

R squared on the single variable runs scored went from 0.32 to 0.39





Many Beautiful Linearly Related Features

Many high correlations (0.5 - 0.9) in correlation matrix of features

Counting Statistics / Career Totals
Runs Scored, Hits, Home Runs, Runs Batted In, Stolen Bases

Rate Stats for Career

Batting Average:	Hits per At Bat
On Base Percentage:	Reaching Base per At Bat
Slugging Percentage:	Total Bases per At Bat

All are “good” stats where higher total or rate is better

In simple linear regression with all variables saw some statistically significant negative coefficients due to the multicollinearity

If you build an algorithmic model

Started with simple , ridge, second and third degree polynomial on full feature set

During cross validation, linear and ridge performed best (~0.35 R2) but second degree polynomial (“Poly2”) had some very high fits on certain folds

Would Regularization and reduced feature set of Batting Average, Home Runs and Stolen Bases help the overfitting Poly2 experienced on certain folds?

Final pipeline

Hold out 20% of data / Cross Validation on remaining 80% for best alpha for Linear and Poly2 / Reshuffle the folds and test Linear and Poly2 with best alpha / Select best model and retrain on full 80% / Test on the 20% held out

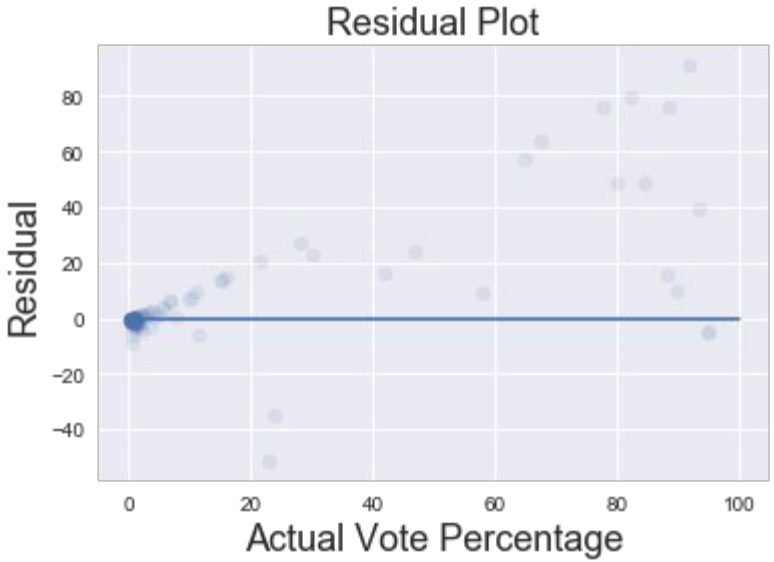
It's not what you want!

Linear Ridge mean cv r^2: 0.303 +- 0.142
Poly2 Ridge mean cv r^2: 0.385 +- 0.135

Final Retrained Poly2 Classifier: 0.518

Root mean squared error: **19 percentage points**

Looked at reducing the data set to the 131 with more than 5% of the vote (players are dropped from ballot if reach less than 5%)



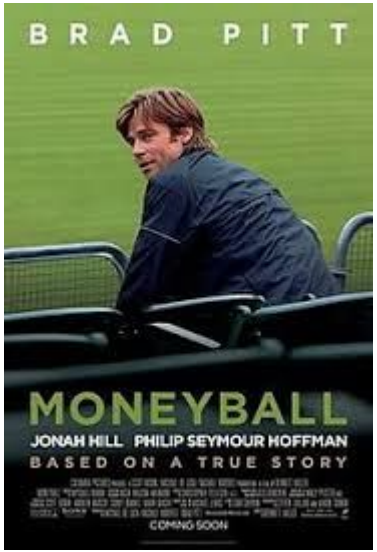
Heads I win, Tails you lose

Has the predictive power of new advanced stats (“Moneyball”) grown over time?

Wins Above Replacement (“WAR”): How many total wins above a replacement level player from all components of player performance (hitting, fielding, baserunning)

Include WAR in full feature set and run LARS Path with Lasso separately on 20th Century and 21st Century

As regularization strength decreased, first variable chosen with Lasso in both time periods was WAR



“It always saddens me to leave the field. Even fielding the final out to win the World Series, deep in the truest part of me, felt like death.”

The Art of Fielding