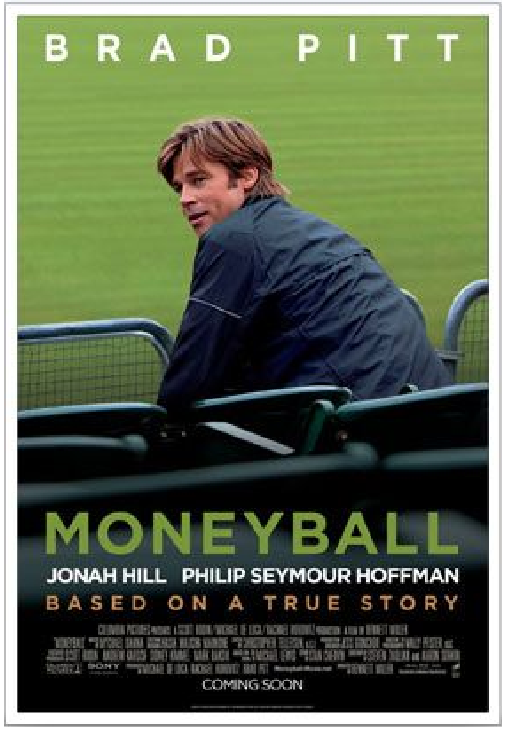
Player Pricing: Using Performance to Predict Salaries in MLB

Ben Stan
General Assembly DS 20 - Winter 2016

Motivation

- Baseball: History of statistics/analytics
- Various means of evaluating players and new metrics always being created
  - WAR (Wins above replacement)
  - DRS (Defensive runs saved)
  - EqA (Equivalent average or BA independent of park)
  - BABIP (Batting average on balls in play)
- **Question**: Is it possible to predict hitter salary based on performance and what are the most important factors? (Prediction + Interpretation)

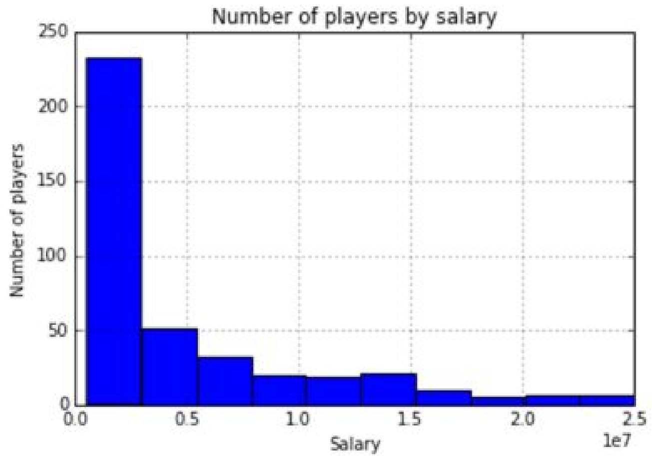


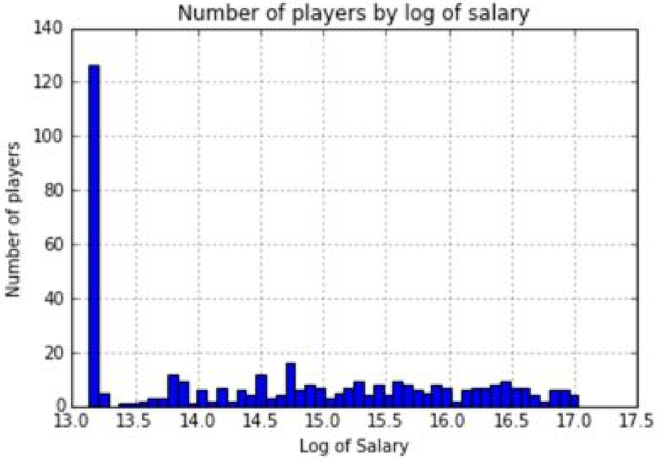
Getting and cleaning data

- Used “The History of Baseball” data set from Kaggle
- Five sources: Batting, Fielding, AllStar, Salary, Player
- Considered player stats in 2014 and salaries in 2015
- Final feature set included
    - Games (g)
    - At bats (ab)
    - Runs (r)
    - Hits (h)
    - Doubles (double)
    - Triples (triple)
    - Home runs (hr)
    - RBI (rbi)
    - Walks (bb)
    - Intentional walks (ibb)
    - Batting average (ba)
    - Slugging percentage (slg)
    - On Base percentage (obp)
    - Stolen bases (sb)
    - Caught stealing (cs)
    - Strikeouts (so)
    - Hit by pitch (hbp)
    - Sacrifice hits (sh)
    - Sacrifice flies (sf)
    - Hit into double plays (g_idp)
    - Number of outs played in field (inn_outs)
    - Put outs (po)
    - Assists (a)
    - Errors (e)
    - Double plays (dp)
    - All Star status (was_all_star)
    - Age (age)
    - League (in_al)
    - Position (pos_)

Working with salary data

- Incomplete data: Batting stats for 1320 players and salary info for only 817
  - 404 observations in final set once pitchers and missing values were removed
- Summary stats
  - Mean: $4.75M
  - St. Dev: $5.78M
  - Min: $508k
  - Median: $2.05M
  - Max: $25.0M
- To remove skew in salary data, took natural log (see graphs) - Removed interpretability from approach





Using linear regression

- Positive initial results: $R^2 = 0.70$
- Most features highly correlated
- Lasso regression too demanding, performed Ridge instead
- Coefficients lacked interpretability
- MSE = 0.60

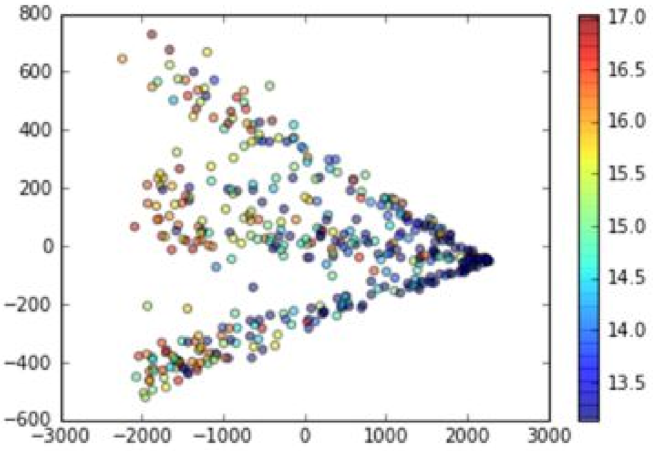
Variable	Coefficient
age	0.16
sf	0.042
g_idp	0.031
bb	0.016
h	0.012
so	0.0046
g	-0.012

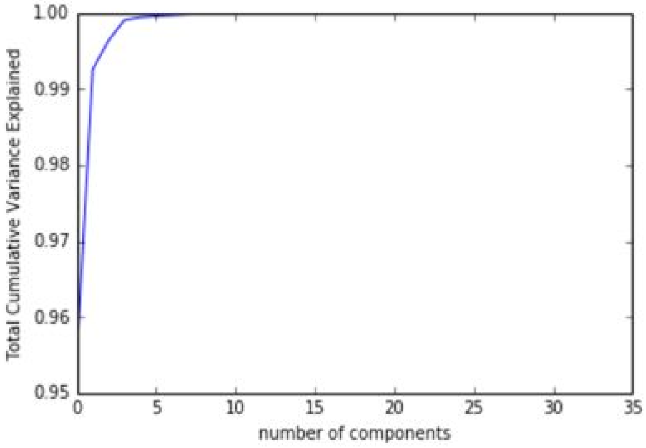
Using decision trees and related approaches

- Used decision tree models to increase predictive power and reduce MSE
- Simple decision tree regressor
    - Depth of 3
    - MSE = 0.69 (Below linear regression)
- Random forest regressor
    - Six features per node
    - 1000 trees per estimate
    - MSE = 0.60
- Gradient boosting regressor
    - Depth of 1 per tree
    - Learning rate of 0.1
    - 100 trees per estimate
    - MSE = 0.56 (Best performance)

Alternative approach: PCA

- Looking to reduce feature set
- Over 99% of variance explained in first two principal components
- Used with K nearest neighbors regressor
  - Five neighbors
  - MSE = 1.28
- Used with simple decision trees
  - Depth of 2
  - MSE = 1.25





Conclusions

- Gradient boosting regressor had strongest performance (MSE = 0.56)
- Feature importances for both advanced tree methods look similar
- PCA captured variance but did not perform as well as tree methods

Rank	Random Forest	Gradient Boosting
1	**age**	**age**
2	**bb**	sh
3	**rbi**	**h**
4	**h**	a
5	**ab**	**bb**
6	double	**rbi**
7	r	**ab**
8	g	**g_idp**
9	**g_idp**	po
10	inn_outs	so

Similar study: Magel et. al

- Similarities
  - Broke apart pitcher and hitter data
  - Took natural log of salary data as well
  - Technique: Stepwise regression was primary method for this publication
- Differences
  - Considered same-year salary (not following year)
  - Only considered players with 400 at bats or 30 innings pitched
  - Created separate model for career stats versus year-to-year performance
- Results
  - MSE = 0.97; R^2 = 0.35
  - Significant features: Total Bases, Total Bases Squared, **Games**, **Sacrifice Hits**, F
    Caught Stealing, **Runs**, **Ground into Double Play**, **At-Bats**, and Stolen Bases

Next steps

- Obtain missing salary information
- Consider players in contract years - Expect improved performance and removes issue of guaranteed salaries in MLB
- Increase interpretability - split data set if necessary
- Perform similar analysis on pitchers