Understanding Career Progression in Baseball Through Machine Learning
Brian Bierig, Jonathan Hollenbeck, Alexander Stroud
(bbierig@stanford.edu, jonoh@stanford.edu, astroud@stanford.edu)
CS229, Fall 2017

Motivation
- Baseball teams retain rights to the first 6 years of service of players they draft. Players typically win the largest contracts after this period, since all teams can bid.
- Contracts frequently exceed $100M in value ($325M record). Even role players often earn a yearly salary of ~$10M. Franchise valuation typically hovers around $1-2 billion.
- Sports literature mainly contains simplistic analyses of aging. We applied ML to predict player value after team control ends.

Dataset
- Main source was Kaggle. We scraped WAR (wins above replacement), the gold-standard player value metric, from Baseball-Reference.com.
- Filtered out all players who started before 1970 and whose careers spanned fewer than 7 years. Also excluded pitcher batting data. This included >80% of contemporary data.
- Joined data sets containing WAR, age, biometric, positional data by year and player. Stacked rows by player so that each training example consisted of one player’s first 6 seasons. Then attempted to predict WAR in subsequent years.
- All analysis completed twice (for batters and for pitchers)

Pre-Processing and Feature Selection
- Binarized categorical features (such as year, fielding position, and handedness).
- Normalized all features using min-max method.
- After stacking 6 seasons of data, rows contain 200+ features for each player.
- Computed additional rate features (like home runs per at bat) from raw data
- Used recursive feature elimination with linear ridge model to select inputs. Asymptotic results generally achieved beyond 15-20 features.

Models
After selecting 15-20 top features, we trained four models. Used 80% (1179/1028 batters/pitchers) of data for training/evaluating models with 3-fold cross-validation. Held out remaining 20% (394/258 batters/pitchers) of data for unseen test set.

Results by Year
- All models perform better than standard lit. approach ("delta")
- Ridge, NN, and SVR perform similarly. Random forest slightly worse.
- Predictions actually improve in later years, as model can infer which players will no longer be active.

Conclusions/Future Work
- Best predictors of future value are cumulative WAR in the first 6 seasons and WAR in 6th season.
- Age factors in more for years 9-11 than years 7-8. Age also more critical for batters (may be that injuries contribute more to pitcher variation).
- Batters in general are easier to forecast than pitchers (R2 of ~0.6 vs ~0.4). This agrees with literature.
- Most features (decade, biometric, rates, position, etc.) not critical to analysis. The two WAR features are sufficient to build a performant model.
- Future work could include sensitivity analysis on which players are included in the model (ie, different min thresholds)

References