CS229 Project Proposal
Peak Baseball Career Performance from Early Career Indicators

Category: Athletics and Sensing Devices

Paavani Dua (paavanid), Liam Kelly (kellylj),
and Matthew Lee (mattskl)

19 October 2018

For our project, we intend to investigate how major league baseball players evolve over their career. In particular, we would like to look at player performance, and predict what season in a player's career they are at their peak based on their first three seasons. This is an application project and this type of analysis is motivated by a team interested in trading for different players and fantasy baseball players choosing teams at the beginning of each season. Knowing which season a player is expected to be at their peak (barring injuries) will provide teams knowledge on when to hire someone and for players to know if they should be asking for more money in a contract. A team looking at a multi-year contract with a player that would include the predicted best year may also consider this analysis when negotiating the contract.

When deciding which sport we wanted to investigate, we settled on baseball due to the large number of games (162 in a season) and relatively low number of injuries providing a large amount of data for us. The project will entail deciding key performance indicators (KPIs) to focus on, such as looking at different positions ie batting, fielding and pitching, the players ages, the team they play for each season, their salary and running linear regressions on them to predict KPIs per season, then computing a score for those to determine the max for each player. By looking at different positions individually for the different KPIs, we anticipate running a number of experiments to evaluate the algorithm and determine which input features are best for each position. Furthermore, we would look at applying SVR which is an extension of an SVM for linear regression to learn KPIs without supervision and employ LSTM since a person's career has a strong time dependence.

While there is numerous previous research into the value of a sports player, many focus solely on the player's career projection [a][b], many focus on immediate performance only [c][d] and make predictions based on their age range [e]. Wins Above Replacement (WAR) is also widely used as a metric to compare players to each other and their contribution to their particular team, but does

not predict a player's growth. We hope to leverage the techniques used in these works to help find players at the highest performance stage of their career.

We plan to use the Baseball Data dataset from Kaggle, which contains statistics on each player every season since 1871. Our first task would be to preprocess the dataset and edit KPIs since age has not been given as a KPI in this particular dataset, then further scrape through the data since statistics tracked have changed overtime. Once we have cleaned this dataset, we will be able to run through the above algorithms and predict peak performance. Specifically, we would set aside a subset of players to exclude from the training set, and see if our trained model produces accurate predictions on this validation set.

[a] Understanding Career Progression in Baseball Through Machine Learning - http://cs229.stanford.edu/proj2017/final-reports/5216878.pdf
[b] Predicting Career Paths of NBA Players - http://cs229.stanford.edu/proj2012/ShahCouslandRobbins-PredictingCareerPathsOfNBAPlayers.pdf
[c] Beating fantasy football - http://cs229.stanford.edu/proj2016/report/Fox-BeatingDailyFantasyFootball-report.pdf
[d] Machine Learning for Daily Fantasy Football Quarterback Selection - http://cs229.stanford.edu/proj2015/111_report.pdf
[e]What is a baseball player's prime age? https://www.bostonglobe.com/sports/2015/01/02/what-baseball-player-prime-age/mS39neFWm4hrVukT6lSYuK/story.html