**Personal Project-**

**Fantasy Baseball (On-going): Developed machine learning model for estimating fantasy baseball performance based on historical Major League Baseball data.**

* Analysis in R using dataset with over 12 million observations and over 200 columns with data from 1952-2017. Analysis focuses on data from 2005 to the present due to fundamental changes in the game that impact offensive performance, and to exclude bulk of steroid era that greatly inflated offensive player performance.
* Data munging using base R aggregate function to transform raw data from plate appearance level to game level to reduce noise.
* Game level data split into four matchup datasets of batters and pitchers based on dominant hand to take advantage of baseball splits.
* Data mining was used for feature engineering of modeling inputs. Used combination of R functions: Data.table, and rollapply to create rolling averages of player performance. Used for loop to create averages of statistics from 1-10 games. Created statistical player performance inputs based on rolling time periods and batter vs pitcher matchups. Input features consist of statistical, categorical, and weather inputs.
* Predictive modeling goal is to predict the probability that player scores some non-zero amount of fantasy points and based on ranking of probabilities, player performance is mapped back to discrete point counts.
* Predictive modeling approach is a stacked ensemble approach combining the results of individual machine learning algorithms: Logistic Regression, Random Forest, Gradient Boosting, and Neural Network. Algorithms are estimated and then combined through the H2o package in R. Optimal lineup is then selected based on predicted points subject to the salary cap constraint.
* Created functions in R for diagnostic testing, testing includes: measuring discriminatory power of model for rank ordering using modified Somers’ D statistic, analysis to quantify for incorrectly ordered pairs how many notches off the prediction is from reality, and the correlation in the predicted point counts with actual counts.
* First generation of model to be implemented for the 2018 season, further analysis and improvement for modeling unknown matchups of batters vs relief pitchers, and developing further diagnostic testing for individual algorithms, ensembled model, and full system modeling approach. Potential weighting will be explored for giving more weight to recent years over older years.