**Predicting NBA Player Performance**

**Team members**

Patrick Miller, Vandana Ramakrishnan, Nathaniel Matare, Ernie Mori, Jordan Bell-Masterson

**Data Set**

We have aggregated 4259 features for a dataset that includes 97,866 observations of individual players appearing in the 2013 - 2016 NBA seasons. The dataset is time series dependent; each observation at time t-0 includes lagged statistics leading up to time t-0, and known information at time t-0.

* Player biographical information known at time t-0
* Game atmospherics at time t-0
* Individual player statistics at time t -1 through time t-10
* The player’s current injury or lack thereof at time t-0 through time t-10
* The players team statistics at time t -1 through time t-10
* The opposing teams statistics at time t -1 through time t-10
* The ‘type’ of player [similarity] as determined by season averages via k-means clustering

Additionally, the dataset includes matchup information encoded as [-1, 0, 1]:

* [1] if the player or team is either the player or on the players team
* [0] if the player, team or official was not present during the game
* [-1] if the player or team was opposing the player

Further, for each player we have aggregated news and analyst reports between games played:

* The probability of the text falling into one of twenty topics; as determined by an LDA model.
* N-gram tokenization of forty specific key words and phrases; reported as frequency.

**Project Idea:**

We will use this dataset to predict the number of points a player will score at time t-0 given information at time t-0 and historical information available until time t-0.

Given the large number of features relative to an individual player observation—at the most ~260 player observations to 4259 features—we will likely employ dimensionality reduction via PCA, MCA, or clustering. Further, because of the time series dependent nature of the dataset, cross validation will be tricky.[[1]](#footnote-1) Thus, we have written several utility functions to perform time-forward cross validation.[[2]](#footnote-2) We hypothesize that we will face a highly interacted environment affected by player idiosyncrasies and game matchups; thus we intend to employ Deep Neural Networks. However, we may find better success with a combination of models and are therefore considering ensemble models: a combination of trees and regularized regressions. Additionally, we may have several sub-prediction problems; that is, predicting win probability may help us better estimate a players point prediction.

Please see the appendix for additional information.

**Appendix:**

1. http://robjhyndman.com/hyndsight/tscv/ [↑](#footnote-ref-1)
2. https://github.com/nmatare [↑](#footnote-ref-2)