National Basketball Association Most Valuable Player Award Prediction using Neural Network

Group Members: **Louis Rabinowitz**, Tom Boyd, Clark Davis, Marco Liu, Nishit Bade

The most prestigious award that can be given to any individual NBA player each season is the Most Valuable Player award, better known as the MVP award. The award is given, as its name suggests, to the player who is perceived as most valuable. We would like to predict the winner of this award using just box score statistics (points, rebounds, assists, blocks, steals, turnovers, minutes played, games played, shooting percentage, etc.) Data for most of these statistics is available for each player starting with the 1950 NBA season on www.basketball-reference.com. Some statistics are not available over the whole time range (ex. Blocks were not recorded until the 1973-1974 NBA season.) Obviously, no linear combination of these statistics will be able to predict the winner of the MVP award with absolute certainty (otherwise Vegas would lose a lot of money.) Therefore, we propose using a multilayered neural network to predict the outcome of the MVP award. Many advanced stats (some combination of box score stats) have been created to evaluate a player’s value. We will be able to compare these to our results to determine the viability of using a neural network to predict the MVP. A more stringent test to validate our results would be comparing them to the Las Vegas odds which have the advantage of taking in more features than just box score stats (also do not necessarily try to reproduce actual odds).

There are a few hurdles that must be crossed prior to the completion of this project. The first has already been mention is the lack of data for some statistics. The easiest solution to this is remove all data from 1950 to 1973. This removes about a third of our data set (actually less because there were fewer players.) Another issue is there will be an overwhelming number of non-MVP winning players each season in comparison to MVP winning players (roughly 500:1). This can be fixed by a combination of over sampling the MVP winners in our training set and over sampling our non-MVPs in our training set as described in [Chawla, 2002](https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/chawla2002.html). The next two largest problems we foresee are the power of narrative in selecting an MVP and the change in playstyle over time. There is no way to correct for narrative using just box score data. We can adjust for change in playstyle by using pace (possessions per game) to standardize our data.