# Predicting Games Played

*Kyle Joecken*

Injuries: The bane of any promising fantasy season.

## Question

Can we use games played data from the previous few National Hockey League (NHL) seasons to predict how many games a particular skater (*i.e.*, non-goalie) will play in an upcoming season?

In particular, the hope is that games played for various players in their more recently played seasons will help to predict how many games they'll play in future seasons. This is quite difficult, as the data are exceedingly noisy due to the largely random nature of injuries in a fast-paced, full-contact sport. Surely there is some small signal for which we can search.

## Data

First, we load the data from our local (scraped) database of season-wide NHL data. It is stored in a data frame called `skaterstats`.

The data was scraped from the NHL.com statistics pages by a Python script using the Scrapy framework; the relevant script and a `CODEBOOK.md` file can be found at this link.

## Features

In order to predict games played in 2014, we'll need to first decide what our predictors are. Let's take all players with games played in each of the 2012, 2014 seasons and use `varimpplot()` on a random forest model to select predictors. We'll also throw away team information (columns 3 - 5) and shootout information (38 - 41), as both are incomplete and the former is unlikely to be relevant.

```
skaterstats <- subset(skaterstats, season > 2011)     # shed old data
skaterstats <- skaterstats[, -c(3:5, 38:41)]          # drop team/SO variables
skaterstats <- reshape(skaterstats, timevar = "season",
                       idvar = "nhl_num", direction = "wide")
skaterstats <- skaterstats[, -(81:118)]               # drop non-GP 2014 vars
```

We'd also like to add a simplified age to the computation (`2014 - birthyear`).

```
library(lubridate)
skaters <- skaters[, c(3,5)]                    # grab nhl_num, birthday
skaters$age <- 2014 - year(skaters$birthday)    # compute "age"
skaters <- skaters[, -2]                        # drop birthday
skaters <- merge(skaters, skaterstats)[, -1]    # merge along/drop nhl_num
```

Finally, we reduce our collection of 1175 NHL skaters (for now) down to the set that played in all three of the last few seasons.

```
data <- skaters[!is.na(skaters$games_played.2012) &
                       !is.na(skaters$games_played.2013) &
                       !is.na(skaters$games_played.2014), ]
```

We build our model on the remaining 611 skaters.

```
library(caret)
library(randomForest)
testMod <- randomForest(games_played.2014 ~ ., data = data, importance = TRUE)
print(testMod)
```

```
##
## Call:
##  randomForest(formula = games_played.2014 ~ ., data = data, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 26
##
##           Mean of squared residuals: 345.6
##                     % Var explained: 39.87
```

```
varImpPlot(testMod)
```