Adjusting NHL Corsi for Quality of Opposition

DP Rooney

The Problem

Hockey has seen a lot of growth in advanced stats in the last few years. There is a family of stats called "Corsi" stats that look at shot attempts (shots on goal + shots missed + blocked shots). Individual Corsi stats (as opposed to team Corsi stats) look at the shot attempts that occur while a given player is on the ice. The so-called "Corsi For Percentage" (abbreviated CF%) is defined as follows, where shot attempts are abbreviated SA:

CF% (player X) = sum_j(SA_j | X's team, X on-ice) / sum_j(SA_j | both teams, X on-ice).

This is seen as a useful metric because it eliminates some of the randomness associated with goal-scoring, and it combines offensive and defensive merit (a gifted offensive player may have lesser value if his defense is atrocious). Additionally, shot-attempts occur more frequently than goals by one or two orders of magnitude, so our data set is considerably bigger.

One major drawback with this stat is that it does not take into account the quality of opposition. A fundamental phenomenon in hockey is line-matching. Each team has four different forward-lines (and three defense-pairings) that generally differ in abilities, and coaches will strategize as how to optimize the match-ups between opposing lines. For example, a good defensive forward line will often be matched against the opposition's best offensive line. Therefore, comparing the CF% of a first-line and a fourth-line player is unfair to the better player, since he has to work harder to generate a shot attempt, and prevent opposing shot attempts.

For example, this reddit post points out that Jake Virtanen, a young, raw prospect for the Vancouver Canucks, has a better CF% than Jonathan Toews, who is seen as one of the best players in the league:

https://www.reddit.com/r/canucks/comments/4omaqp/why_jake_virtanen_is_better_than_jonathan_toews/

While Virtanen may develop into one of the best players in the league, it is unlikely that his performance was on par with Toews' last year.

The goal of this project is to develop a model that is more nuanced. Instead of modeling the ratio of shot-attempts as a function of a single player, we will consider the probability of shot-attempts as a function of all players on the ice. In other words, we want to model the probability

P(next SA is for home-team | X_1, … X_6 are on-ice for home-team,
Y_1, …, Y_6) are on-ice for away-team)

The Client

Professional sports is a lucrative business, and while hockey is less prominent than some others, there is a strong financial incentive to find and exploit market inefficiencies. Hockey is also a sport sometimes clouded by clichés, and savvy upper management would like to pierce through the common wisdom if data-based methods allow it. So our ideal client would be a professional NHL franchise. In particular, a team may consult a data scientist prior to the off-season free agency period. If there are certain players available, prospective teams would like to know which of them deserve a shiny new contract. A predictive model that describes shot-attempt probability relative to opposing players would allow a team to directly compare players and assess their relative value.

The Data

The NHL publically releases game-by-game data online. Here is an example of an individual game report, and here is the roster report for the same game. Salary information is available here. There are scrapers for the game-by-game data available on-line in both R and Python. The R-scraper however does not appear to scrape individual shot attempts, and the Python scraper is only available as a pre-release (and appears to be buggy). Therefore, we will likely have to code our own scraper. Nevertheless, we should be able to obtain a healthy data set. We estimate the total number of shot-attempts in a given season to be between 100,000 and 200,000.

The Approach

The base model that we want to use is a logistic regression. Assign a coefficient to every player, then, to calculate shot-attempt probability, sum the coefficients of the home players, subtract those of the away players, and take the logistic function. One can add a constant term to model home-ice advantage. We will train the regression on a training set of shot-attempts, and measure the cross-entropy of the algorithm on a cross-validation set. This will be compared to a cruder algorithm that averages the CF% of on-ice players and uses this to assess probability of shot-attempts.

Ideally, we would like to also consider more sophisticated models:

- neural networks
- random forests
- Deep Learning

Progress on the logistic regression will dictate further modelling. At the very least we would like to train a logistic regression, and one neural network.

The Deliverables

1. Python code, for the scraper, for training our algorithms, and for interpreting our results
2. A .pdf report
3. Slide presentation