Capstone 2

Based on historical regular season performance of their respective players, which NHL teams will qualify for the playoffs in the 2022 and 2023 NHL seasons?

Premise

The NHL analytics revolution is well underway. Many NHL franchises now employ a full analytics department and others are just beginning to build out their data teams. An NHL data scientist's role is especially complex because hockey is more likely to be influenced by random events than most other major sports. For example, NHL players have short shifts and limited ice time, meaning a "great" NHL player has less time to influence a game than, say, an NFL quarterback who plays every offensive snap for his team. Hockey games also tend to have fewer scoring events than other major sports, so an unexpected bounce of the puck (a common occurrence) can be the difference between a win and a loss. An NHL data scientist must therefore pull meaningful trends from incredibly noisy data.

Criteria

The overall aim of my project is to predict the NHL teams that will make the playoffs. My threshold for success is 75% precision - I want to predict at least 12 of the 16 qualifying teams each season. To do this, I'll likely employ more than one model. My first model will assess individual players in an effort to predict each player's performance for the upcoming season. The second model will utilize those predicted stats for the unknown season to make its prediction about team performance.

Data and Scope

The data for this project will be sourced from hockey-reference.com, starting with the 2005-2006 NHL season. This season marks the beginning of the NHL's salary cap era. I'll use time-series cross-validation to train the models progressively. The NHL's playoff criteria changed from top-8 conference teams to top-3 division teams plus 2 conference wildcards in 2013-14. I may have to account for this, depending on how I choose to model team success.

Constraints

Other constraints include limited team history (Seattle and Vegas), changing rosters (trades, acquisitions, call-ups, and injuries), and limited player history (rookies). I am eager to see how the models perform in spite of these constraints. The trade deadline in the NHL falls relatively late in the season, so it can be argued that deadline acquisitions are more likely to influence playoff success than regular season success. If my model is unable to meet my desired metrics for success, I may explore the influence of trades more deeply.

Stakeholders

I intend to reach out to data professionals working in the NHL for advice as I tackle this project. Though simplified, this is the type of model that an NHL team would use to assess the strength of their roster and find weaknesses. As the sports gambling movement continues to expand, bettors would likely find tons of value in this type of model as well.

Future Considerations

A common method of predicting season success is to simulate each NHL game for the entire season a number of times and let the most likely outcome stand. I may choose to explore this route in the future, but for the sake of this project, I'd like to instead use a classifier to simply evaluate the strength of each roster against the others and place teams into "playoffs" and "non-playoffs" categories.

Summary

NHL fans can generally predict most of the top-10 teams before a given season, but there are always some surprises. I'm eager to see if this simplified, model-based approach can still achieve reasonable precision, and I'll compare its performance on the test data against archived pre-season predictions from fans and NHL pundits.