NHL Stanley Cup Predictions

Erdos Institute Data Science Bootcamp 2022

Superstars:

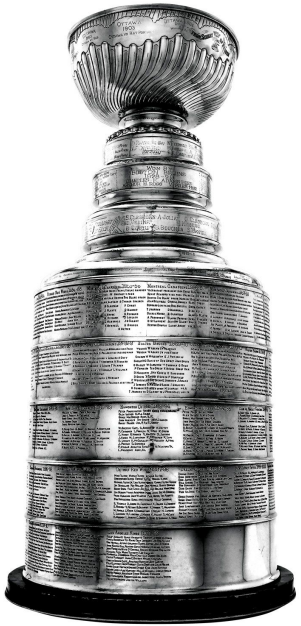
Dylan Bates Chenyi Gu Kanishk Jain Briana Stanfield

Introduction and Problem Statement

Hockey's Stanley Cup is the oldest existing trophy to be awarded to a professional sports franchise in North America, and the often considered "**the hardest trophy to win in professional sport**".

Using just regular season data, we want to answer the following:

- Which features predict playoff success?
- What influence do players have on team outcome?
- Ultimately, who will win the Stanley Cup?



Data Collection

We collected a wealth of **player** and **team**-level data on wins, goals, shots, hits, blocks, penalties, etc. from `NHL.com` and `hockey-reference.com` going back to the 2005-06* season. Data collection methods included:

1. Directly scrape from websites to get team-level data for each season;
2. Using NHL API Python wrapper: `nhlpy`. For the player data, we pulled the roster from each game, and calculated a weighted average of stats at each position. Weights were based of the mean time on ice per player.

*There were significant rule changes after the 2004-05 lockout, including removing the possibility of a tie game. Data from before this was also missing important features that could not be easily inferred.

Exploratory Data Analysis

- Predictive **team** features:

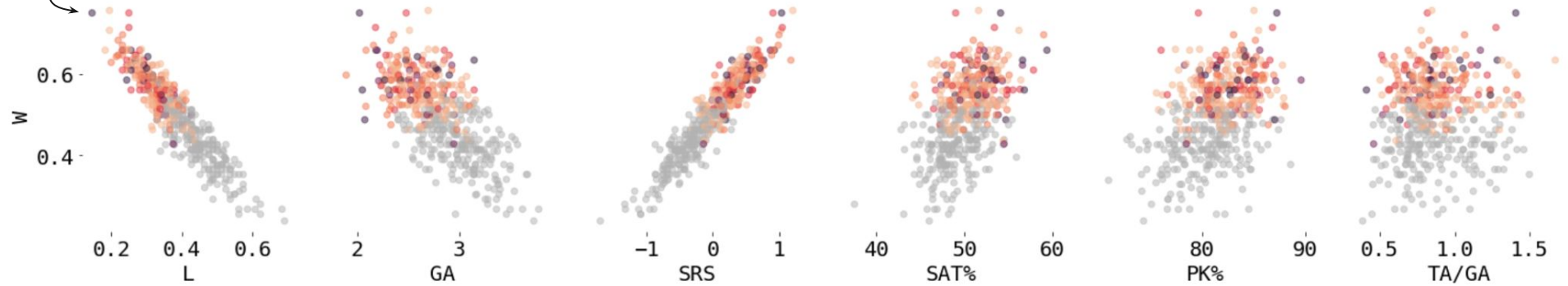
  Wins (**W**), Losses (**L**), Simple Rating System (**SRS**), Goals Against (**GA**), Corsi (**SAT%**), Penalty Kill (**PK%**), and the calculated feature Takeaways/Giveaways (**TA/GA**).

- Predictive **player** statistics:

  Plus/minus (**+/-**), goals (**G**), and assists (**A**) were the most predictive of wins.

  Penalty infraction minutes (**PM**), hits and blocks had a slightly predictive element

darker means making it farther in the playoffs



Predictability of Hockey Results

- Luck makes up 38%* of winning a hockey game
- This is impossible to model without training on test data.
- We can model the uncertainty with a Bayesian model, predicting the probability of winning a game

*_Forecasting Success in the National Hockey League using In-Game Statistics and Textual Data_, Weissbock 2014

Modeling Approach

We modeled the data with 4 methods: Neural Network, Logistic Regression, AdaBoost, and Random Forest. We also created an ensemble of the 4 models.

Method	Train accuracy	Val. accuracy	Test accuracy	Playoff Winner
Neural Network	71.38%	70.22%	70.01%	COL, MIN, CGY, STL
Log. Regression	62.24%	62.82%	61.64%	Calgary Flames
AdaBoost	62.29%	62.09%	56.16%	Toronto Maple Leafs
Random Forest	62.93%	61.55%	54.18%	Calgary Flames

Although each was only able to predict wins with <72% accuracy on the training set, they were able to consistently predict playoff wins with >62% accuracy.

Results

Accounting for our predictions and the sport's uncertainties, this year's trophy would go to….

BUT...

They only have a 17% of winning. If they get eliminated early on, then the Calgary Flames (14%) or St. Louis Blues (13%) could win too.

Predictions as of June 4, 2022

Colorado Avalanche: 34.6%

Edmonton Oilers: 3.8%

New York Rangers: 42.0%

Tampa Bay Lightning: 19.6%

Conclusions

- Hockey is a difficult sport to predict, with underdogs winning far more than in any other professional sport;
- We can update the probabilities as the playoffs go on, accounting for these upsets;
- Our model **performs better** than comparable models that do not include roster data.