TEI Hockey Predictions **Team Skylab: Spring 2022**

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

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

- Which features predict playoff success?
- What influence does individual player performance have on the outcome of the game?
- Ultimately, who will win the Stanley Cup?

This data is useful to any hockey fan, including gamblers, casinos, and the media.

Data Collection

We collected a wealth of player and team-level data on wins, goals, shots, hits, blocks, penalties, etc. from going back to the 2005-06 season to 2021-22 season. Data collection method included:

1. Directly scrape from hockey-reference.com and nhl.com;
2. Using NHL API Python wrapper: nhlpy

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. There were 62 total features: 62 (47 features from team-level data, 15 from player-level data). The most important features in predicting playoff success are Wins (W), Losses (L), Simple Rating System (SRS), Goals Against (GA), Corsi (SAT%), Penalty Kill (PK%), and the calculated feature Takeaways/Giveaways (TA/GA). For the player data, plus/minus (+/-), goals, and assists were most predictive of wins. Penalty infraction minutes, hits, and blocks were also slightly predictive

Modeling Approach

For our final model, we collected the roster from each game and created a weighted average of player statistics at each position (forward, defenseman, and goalie). Weights were based on the average time on ice of each player in a given season. This data, along with the home and away team statistics were fit to four separate models to predict which team will win any given matchup. For all models, the training and validation set were from the regular season, and the test set was from the playoff data.

Sklearn classifiers: We add Logistic Regression, Adaboost and Random Forest.

TensorFlow: We are using TensorFlow Keras Sequential Neural Network. Hyperparameters are optimization algorithms, loss function, layer structure of NN, learning rate, number of epochs, batch size, dropout rate.

Overall, we could consistently model a game win with a 68.89% accuracy using our neural network and decided on an ensemble model of all four models as the final model to predict game wins in the playoffs, while reducing overfitting. Knowing the teams who made the playoffs and their matchups, we calculated the likelihood for each team to go onto win the Stanley cup. Our predicted winner for the 2021-2022 season is the Colorado Avalanche with a 16.9% probability of winning. In actuality, the Colorado Avalanche are leading 2-0 against the Edmonton Oilers in the semifinals.

Conclusions and Future Directions

Hockey is an inherently difficult sport to predict. It is noisy and unpredictable, with underdogs winning far more than in any other professional sport. It's been reported that luck makes up 38% of winning a hockey game, which is impossible to model without training on test (playoff) data. Our model was able to account for nearly all of predictability, correctly predicting winning teams up to 70% of the time, and producing a realistic Bayesian model (like the weather) that accounts for this. By using data from individual players instead of aggregated team data, we are better able to account for recent player behavior and roster changes from injuries and trades than existing models.

Because we used regular-season data to predict playoff results, there were some factors unique to playoff games that were not accounted for, and test accuracy was lower than validation accuracy in all models. For example, the expected time on ice for a given player may change in the playoffs based on their performance as a strategic decision. In the future, we could possibly estimate changes in parameters from the regular season to the playoffs and use that to improve model accuracy.