UTAH HOCKEY CLUB

Predicting Individual Skater Playoff Performance with Machine Learning

Utah Hockey Club Analytics Challenge

Andrew Kratsios

August 23, 2024

Overview:

This analysis demonstrates that future playoff performance for an individual skater can be predicted from that skater's historical data. While on the surface the premise seems obvious — past results predict future performance — the nuances between breaking down a superstar's impact on a game versus a third or fourth-line role player under one umbrella becomes an interesting problem to tackle. Immediately questions arise: How do you fairly measure a player's impact on a game? Can a model predict why some players can elevate their game come playoff time, while others shy away from the spotlight? Given the constraints of the salary cap, who are the hidden gems in the league that can perform come playoff time?

To tackle this problem, I created an XGBoost regression model trained on historical data of NHL players from 2014-2023. This data was randomly split into training and testing cohorts to validate the model. Throughout this report, I will walk you through my methodology and results, as well as a current use case for the model. This model can have a material impact on how an NHL general manager and their staff build a Stanley Cup-contending roster. When combined with management expertise, it will help create a team that is ready to compete in the playoffs, not just get there.

Data Sources:

For this analysis, I used data from MoneyPuck, PuckPedia, and the NHL's API. MoneyPuck has historical data from 2008-present which contains player, team, and shot data. For this project, I focused my efforts on the player data. MoneyPuck player data is segmented by the regular season and playoffs and has breakdowns for different game situations (5-on-5, 5-on-4, all, etc.). MoneyPuck also contains player biographical data such as nationality, height, weight, etc. NHL API data was used to build out rosters to score the current NHL landscape with the model. Finally, I used PuckPedia to gather player contract data including salary and UFA status.

Methodology

Problem Formulation:

To define the problem I first determined the sample population timeframe. To train the model I used the 2014-2023 NHL seasons. This is a good time frame because it uses enough historical data to get a large sample size, but also includes recent seasons that give a good indication for future seasons. Next, I defined the action date; when should the model be trained and scored? The action date I chose is the last day of the season before the one being scored. I chose this date because the model should be useable in an actual context: for free agency. To summarize, this model will predict a player's impact on next season's playoffs using past seasons' data at the start of NHL free agency.

Target Variable:

The target variable I chose is inspired by Dom Luszczyszyn's Game Score. Game Score is designed to be a "standardized measurement for single-game productivity." It is a linear combination of goals, assists (primary and secondary), shots on goals, penalties, and more. To fairly judge a skater's impact on a game, normalized for their usage, I created a game score per second on ice variable for a player in all game situations. This final rating is a scaled value between 0 and 1000. Below is a sanity check showing the top players by their mean game score per time on ice (GS_TOI) in the playoffs from 2014-2023.

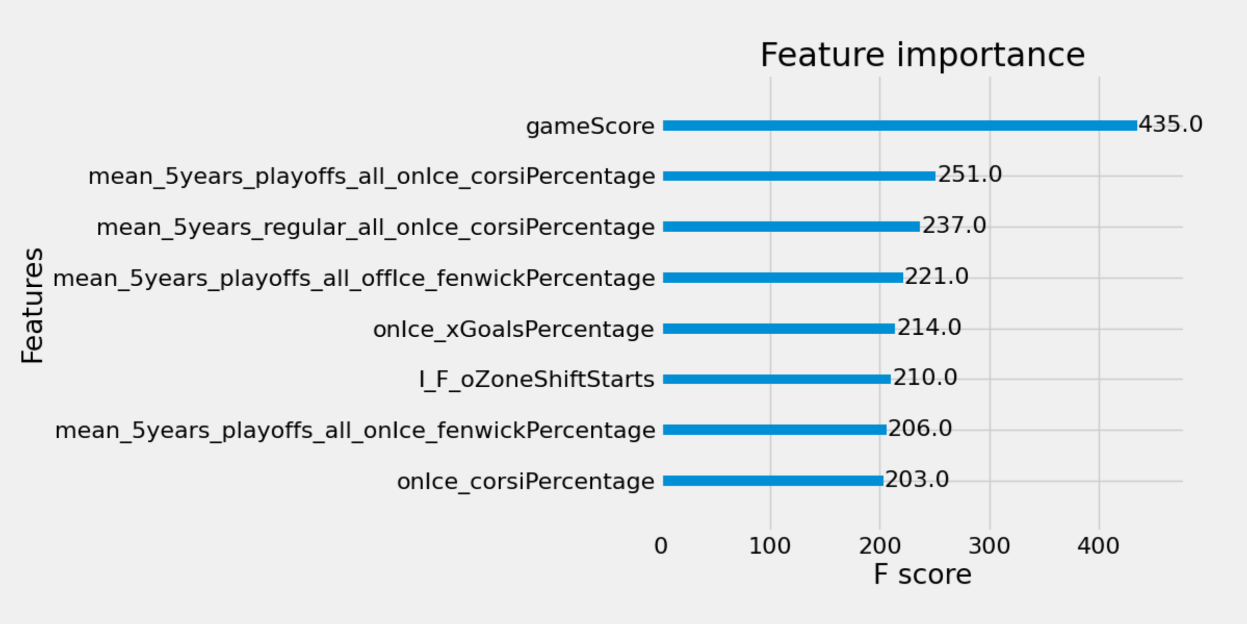
Player	Mean GS_TOI
Nathan MacKinnon	733.14
Daniel Sedin	730.00
Connor McDavid	727.67
Henrik Sedin	714.00
Mikko Rantanen	711.71
Leon Draisaitl	709.00
Nikita Kucherov	704.50

Feature Selection and Engineering:

For the model features, I used a combination of a player's previous regular season and playoff statistics. I have included the historical regular season stats (simple and advanced) such as goals, assists, game score, etc. For both regular season and playoffs, the mean value for each of the past five seasons was calculated for a player. Finally, player biographical data (e.g. weight, height, age, etc.) was included. Unfiltered, this feature list has about 500 different variables and the dataset has about 3000 rows. 500 features are too many for a dataset of this size, so I filtered down the feature list to the top 75 most impactful features on the model's performance using each feature's mutual info (dependence on the independent variable).

Model Training:

The final model is an XGBoost regression model. It was trained using a train-test-split of 80% training data and 20% test data for model validation. This chart shows the top features ranked by importance to the model. The previous season's game score and the historical average Corsi and Fenwick also have a measurable impact on the model.



Model Validation:

To validate the model the root mean squared error (RMSE) for the train and test batches can be compared. The RMSE measures the difference between the actual and the model's predicted values. This is a good measurement for regression models because RMSE is on the same scale as the target variable (0-1000) so it is intuitively understandable (lower is better). For our model, the training data had an RMSE of 6.63828 and the test data had an RMSE of 7.35246. These are great values; not only are both very low, but their magnitude is very similar. This indicates the model is a good predictor of the target variable and it is not overfitting to the training data. Further model validation techniques can be used, but for this model, RMSE coupled with sanity checks for known players on the scored data is enough.

2024-2025 NHL Season Projected Top Playoff Performers:

One use case for this model is to test it on current NHL players to see who might be the most impactful come playoff time next April. As one would expect, dominating the top of the list are proven playoff performers and bonafide superstars in the league. However, one thing we can do is use players' cap hits to see what contracts have the best (and worst) expected value come playoff time. This part of the analysis should be taken with a grain of salt, due to stars in the league making proportionally much larger salaries. However, it does raise the question, in such a team-oriented sport like hockey, what is the optimal proportion of the salary cap to allocate to stars versus the rest of the roster?

Top Players by Expected GS_TOI

Player	Expected GS_TOI
Leon Draisaitl	739.41
Connor McDavid	726.23
Nikita Kucherov	711.38
Mikko Rantanen	700.15
Nathan MacKinnon	694.31

Bottom Players by Expected GS_TOI by Salary Cap Hit

Player	Expected GS_TOI/Cap hit
John Tavares	0.000056
William Nylander	0.000056
Nathan MacKinnon	0.000055
Drew Doughty	0.000053
Auston Matthews	0.000051

Team Level Projections:

The same exercise can be done on a team level. These charts show the model's expected top teams in the 2025 NHL playoffs based on individual player value.

Top teams by expected GS_TOI

Team	Expected GS_TOI
EDM	601.65
FLA	599.45
CAR	597.75
TBL	595.39
COL	593.62

Top Teams by Expected GS_TOI by Salary Cap Hit

Team	Expected GS_TOI/Cap hit
WPG	0.007402
VAN	0.006946
NSH	0.005970
PIT	0.005803
NYR	0.005783

2025 Free Agent Class (who to target for trade deadline):

One practical use case for this model is finding players who can bolster your lineup at the trade deadline. Below are the top 2025 UFAs to watch throughout the season based on their projected playoff GS_TOI and current cap hit. While the big names are much harder to trade for, when sorted by predicted score relative to cap hit, we can create a list of more practical trade deadline acquisition prospects.

Top 2025 UFAs by Expected GS_TOI

Player	Expected GS_TOI
Leon Draisaitl	739.41
Mikko Rantanen	700.15
Sidney Crosby	664.41
Sam Bennett	649.44
Carter Verhaeghe	646.66

Top 2025 by Expected GS_TOI by Salary Cap Hit

Player	Expected GS_TOI/Cap hit
Adam Gaudette	0.000771
Philippe Myers	0.000746
Axel Jonsson-Fjallby	0.000742
Devin Shore	0.000728
Michael Eyssimont	0.000726

Caveats:

While this model is a good benchmark for predicting future playoff performance, there is room to improve. By only using historical NHL statistics, we get a great prediction of how a career NHLer will perform, but fall short on predicting young players and players who have come from the AHL or Europe. Depending on your use case, incorporating this data into the model could be very beneficial. Similarly, there are many more data points relating to a player that can be added to the model to boost its performance. Some examples are shot location, special teams usage, and quality of teammates. Future iterations of this model could include new data sources, experiments with different machine learning models, and more advanced feature engineering and selection techniques.

Conclusion:

When used in conjunction with managerial experience, this model can help strengthen the makeup of an NHL team. Using a combination of a player’s previous performance in the regular season and playoffs, this model accurately and repeatably predicts how a player will perform in the upcoming playoffs. Hockey’s ultimate prize is the Stanley Cup and this model helps to build a team for that purpose.

Resources:

Github Repository: https://github.com/akratsios/hockey_analytics

Luszczyszyn, Dom. "Measuring Single Game Productivity: An Introduction to Game Score." *Hockey Graphs*, 13 July 2016, hockey-graphs.com/2016/07/13/measuring-single-game-productivity-an-introduction-to-game-score/.

"MoneyPuck Data." *MoneyPuck.Com -Download Data*, moneypuck.com/data.htm. Accessed 23 Aug. 2024.

"NHL Player Dashboard." *PuckPedia*, dashboard.puckpedia.com/. Accessed 23 Aug. 2024.

Zmalski. "Zmalski/NHL-API-Reference: Unofficial Reference for the NHL API Endpoints." *GitHub*, github.com/Zmalski/NHL-API-Reference?tab=readme-ov-file#get-team-roster-as-of-now. Accessed 23 Aug. 2024.