Building a Stanley Cup Winning Roster Using a K-Nearest Neighbors Algorithm

Lillian Stensland
Master of Science Data Science
University of Colorado - Boulder
Boulder, Colorado USA

ABSTRACT

The goal of every NHL general manager is to construct a Stanley Cup winning roster. However, due to the limited availability of analytics tools, this task is extremely complex. In an attempt to simplify this process, a k-nearest neighbors algorithm was created, classifying players as one of four line-type players, using publicly available player evaluation statistics. The algorithm was modified to classify players as first liners, second liners, or bottom sixers, resulting in increased accuracy. Finally, the algorithm was modified to classify players as either top or bottom six players, leading to much higher accuracy. With further development, this algorithm opens the doors for increased confidence in player trades, drafting, and contract negotiations.

1 Introduction

In NHL hockey, the ultimate goal is to win the Stanley Cup. After an 82 game regular season, 16 of the total 32 teams make the playoffs and go on to play in a bracket of 7-game series, with the eventual Stanley Cup winner playing anywhere between 16 and 28 additional games. Add in a hard limit on total player contract sizes (called "the cap"), and the Stanley Cup is one of the most difficult trophies to win in sports.

Building a championship-level roster requires balancing player skill, style, and expenses to maximize match-up potential. A roster is made up of 23 players, normally consisting of 12 forwards, 6 defensemen, 2 goaltenders, and 3 "healthy scratches," who do not suit up with the team. The 12 forwards are split into 4 "lines," with the best players making up the first line. This format allows for a larger amount of available cap space to be spent on salaries for players who make up the first line than players who make up the fourth line.

Because roster-building is so difficult, analytics can be a useful tool in player evaluation. However, hockey has invested in data science much less than other sports, leading to a limited number of statistical evaluation tools. The purpose of this project is to build an algorithm that classifies a player into their optimal line using existing player evaluation statistics.

A tool such as this can be used in many ways. A team can run their own players through the algorithm to determine their worth when approaching contract negotiations. A team can run the players from other teams through the algorithm to determine potential trade targets. With modification, this tool could also be used to help scouts evaluate draft prospects.

The following report introduces related developments in hockey analytics, displays the current work towards creating and evaluating the model, and evaluates the potential for future work.

2 Related Work

Due to the free-flowing nature of the sport, data analytics is very challenging in hockey. While some sports, such as baseball, have been revolutionized by the introduction of statistics, hockey has lagged behind. Before the summer of 2014, only 15 data experts worked in the NHL [2]. Over the next 12 years, most teams began to invest in statistical evaluation and the NHL began contracting independent data scientists to advance the game. One of the earliest proponents of hockey analytics was Eric Tulsky, now the general manager of the Carolina Hurricanes [2]. By prioritizing data analytics in his management of the team, Tulsky was able to lead the Hurricanes to great playoff success, including a Stanley Cup in the 2025-2026 season.

The review article “Historical Perspectives and Current Directions in Hockey Analytics” (2019) contains a comprehensive overview of the development in statistical evaluation by experts in the sport. For decades, scouts and evaluators relied simply on the box score and their own “eye test” to determine the potential of hockey players. By 2009, however, data scientists developed methods to measure the quality of competition, to quantify offensive and defensive player deployment, and to determine player performance relative to teammates [1].

Through the early 2010s, analysts began implementing regression models to find players with greater impact on goal scoring and defense. In 2014, the “Corsi” statistic was invented, which combines a player’s shots on goal, missed shots, and blocked shots to identify which players generate scoring opportunities [1]. In the following years, metrics to measure shot and event quality have been developed, to gain a better ability to measure team and player performance. Researchers have developed various expected goal models, which take into account shot quality and player ability to show the story of a game that might not be reflected in the box score [1].

This project uses many of these developed statistics as features in player style and talent evaluation. For example, a modified Corsi statistic, and its successor the “Fenwick” statistic, for each player are included in the datasets being fed into the model. Using these created statistics, this project groups players based on talent and play style to find their optimal line classifications.

3 Continued Work

3.1 Dataset Creation

	Feature	F-Score	p-value
10	I_F_takeaways	12.848551	2.626479e-07
6	I_F_points	12.714543	3.043559e-07
3	I_F_xGoals	11.491071	1.187957e-06
7	I_F_goals	10.312484	4.536448e-06
12	I_F_highDangerShots	10.211760	5.093429e-06
4	I_F_shotsOnGoal	9.360785	1.366126e-05
5	I_F_shotAttempts	9.261434	1.534378e-05
11	I_F_giveaways	7.954008	7.208519e-05
8	I_F_rebounds	7.333552	1.520375e-04
1	onIce_corsiPercentage	5.305445	1.835214e-03
2	onIce_fenwickPercentage	5.252279	1.960956e-03
0	onIce_xGoalsPercentage	5.140822	2.253551e-03
9	I_F_hits	3.276315	2.359990e-02

Figure 1: ANOVA F-Test results for the Stanley Cup winning players dataset

Data for this project was sourced from MoneyPuck.com [3]. The project required a combination of season level data and skater specific data. The first task for this project was to create a dataset of Stanley Cup winning players, containing their line classifications and aggregated performance analytics. For each Stanley Cup winning team in the last ten years, the top twelve forwards by ice-time made up the four lines. For each player, the seasons after they won the Cup were not considered, and the performance analytics were averaged so each player had the following 13 features:

1. **I_F_xGoals**: Expected Goals
2. **I_F_shotsOnGoal**: Shots on goal
3. **I_F_shotAttempt**: Shot attempts
4. **I_F_points**: Goals + Assists
5. **I_F_goals**: Goals
6. **I_F_rebounds**: Rebound shot attempts
7. **I_F_hits**: Number of hits a player has given
8. **I_F_takeaways**: Number of takeaways the player has taken from opponents

9. **I_F_giveaways**: Number of giveaways the player has given to other team
10. **I_F_highDangerShots**: High Danger Shots (Higher than 20% xGoal Value)
11. **onIce_xGoalsPercentage**: On Ice xGoals For / (On Ice xGoals For + On Ice xGoals Against)
12. **onIce_corsiPercentage**: On Ice Shot Attempts For / (On Ice Shot Attempts For + On Ice Shot Attempts Against)
13. **onIce_fenwickPercentage**: On Ice Unblocked Shot Attempts For / (On Ice Unblocked Shot Attempts For + On Ice Unblocked Shot Attempts Against)

An ANOVA F-Test was performed on the dataset to ensure that all features were statistically significant for this classification project, as seen in Figure 1. The top 2 features, I_F_takeaways and I_F_points, were plotted with only players on the first or fourth lines, seen in Figure 2, to display a solid clustering of both groups. However, the two features struggled to differentiate between second and third line players.

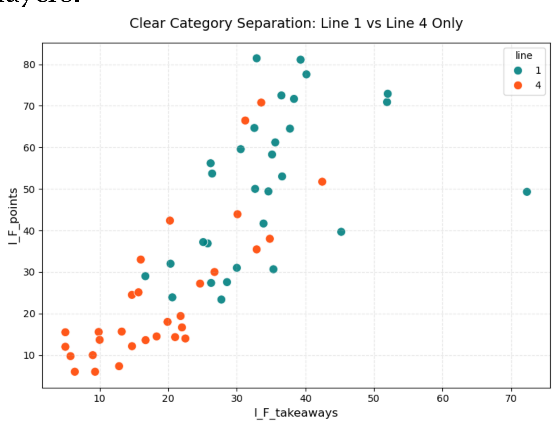

Figure 2: First and fourth line players plotted according to their I_F_points and I_F_takeaways features

3.2 Algorithm Creation

The created dataset of players was then used to create a k-nearest neighbors algorithm that predicted the lineup position of a certain player. By using Stanley Cup winners to train the algorithm, the predictions show a player's optimal position in the lineup, and were not skewed by bad teams with players in positions above their actual skillsets. The algorithm saw an overall accuracy of 41.67%, with a tendency to overpredict players as either first or third liners.

In an attempt to increase the accuracy, the third and fourth lines were combined into a "bottom six" class. This change improved the model's performance, as the accuracy rose to 50.00%. Additionally, combining first and second line players into a "top-six" group improved the algorithm even further. In this case, the accuracy rose further to 66.67%.

3.2.1 Player Examples

Alex Ovechkin, by ice time, was a second line player in the 2025 to 2026 season. However, all three versions of the algorithm predicted that Ovechkin should be a first line player. While Ovechkin took a step back in his playing ability this past NHL season, largely due to age, he is well regarded as one of the best NHL players of all time. Because the algorithm considers Ovi's stats from his entire, highly-awarded career, it predicts him to be a guaranteed first line player. This signals that the algorithm can correctly classify an exceptional player, but is not able to account for age regression.

Ilya Mekheyev, by ice time, was a first line player in the 2025 to 2026 season. All three versions of the algorithm predicted that Mekheyev should be a third line/bottom six player. Mekheyev plays for the Chicago Blackhawks, who were the second worst team in the NHL, a situation which forced Mekheyev to play higher in the lineup than he should. However, if Mekheyev was traded to a Stanley Cup-caliber team, he would be moved much further down in the lineup.

Vladimir Tarasenko, by ice time, was a third line player in the 2025 to 2026 season. The first algorithm predicted Tarasenko to be a second line player, while the other two algorithms predicted him to be a bottom-six player. As Tarasenko played for the Minnesota Wild, who were a Stanley Cup contending team, the algorithm should predict a line similar to his actual usage, as it did. Tarasenko is a solid middle six player, which is reflected by the algorithm.

4 Evaluation

4.1 Near Winners

There are two ways in which this k-nearest neighbors algorithm can be evaluated. The first method is to use *near winners*. In theory, the roster of a team that nearly won the Stanley Cup in a specific season (that is, the season's runner up) should be constructed almost as optimally as the actual winner. By running these rosters through the algorithm and comparing the results to the actual lineups, the efficacy of the model can be determined. The dataset for this method was created by taking the rosters of the last 5 Stanley Cup runner ups:

1. 2021-2022 Tampa Bay Lightning
2. 2022-2023 Florida Panthers
3. 2023-2024 Edmonton Oilers
4. 2024-2025 Edmonton Oilers
5. 2025-2026 Vegas Golden Knights

This dataset saw lower accuracy than the test set used during algorithm creation. Algorithm 1 (predicting 1st, 2nd, 3rd, and 4th line players) saw a 35.00% accuracy and struggled to classify the third and fourth lines. Algorithm 2 (predicting the 1st and 2nd lines along with a bottom six class) saw a 41.67% accuracy and predicted too many players to be first liners. Algorithm 3 (predicting top and bottom six players) saw an accuracy of 65.00% while generally overestimating player talent.

4.2 Changed Winners

The second method of evaluation would require finding players whose line classification changed immediately before winning the Stanley Cup. If a player's career analytics before the Cup win are run through the algorithm and the result is found to match the player's actual lineup position when they won the Cup, the model can be considered effective. The dataset for this method was created by finding applicable players from any point between 2019 and 2026.

This dataset saw much higher accuracy than the near winners evaluation set. Algorithm 1 saw a 62.07% accuracy and struggled to classify the third liners. Algorithm 2 saw a 79.31% accuracy and predicted too many players to be first liners. Algorithm 3 saw an accuracy of 89.66% and had very few misclassifications.

5 Discussion

There was a significant problem faced in the development of this model. In hockey, forward lines are rarely defined by four distinct qualities. Instead, the top two lines and the bottom two lines each tend to be made up of players of similar skill levels and types, with the top-six an offensively minded group and the bottom-six more defensively minded. This resulted in the classification algorithm mislabeling players who belong to the middle two lines. In order to rectify this problem, an algorithm which creates a "bottom-six" group was formed and an algorithm that creates both a "top-six" and "bottom-six" was introduced.

There were seven steps needed to complete this project:

1. Project Proposal
2. Create the original dataset
3. Create the k-nearest neighbors algorithm
4. Create the evaluation datasets
5. Run the datasets through the algorithm and create the evaluation metric
6. Write the final report
7. Complete peer reviews

At the completion of this project, all but the final step have been completed. Peer reviews are an incredible learning opportunity and a chance to explore other focuses of data science. These reviews will be completed in the final days of this course.

6 Conclusion

This report explained the work on a potential hockey player evaluation model, which takes publicly available player analytics and determines the player's line classification. Three algorithms were created, which saw increased accuracy as the specificity in classification decreased. There is much possible work to be done in the future. The original dataset could be refined or modified with different player evaluation statistics, or even with the creation of new ones, to increase the accuracy of the model. The model could also be expanded to include player statistics from before their NHL careers, to allow for modeling of NHL prospects.

As this tool is refined and expanded, it could become very useful to NHL teams. They can use it to determine the values of their current rostered players, to determine where each player should be placed in the lineup or if they are superfluous and should be traded. NHL teams can also use this to determine how much a player is worth, driving contract negotiations. Teams can also use this to determine trade targets on other teams, to fill holes that exist in their lineups.

REFERENCES

1. Namita Nandakumar and Shane T. Jensen. 2019. Historical Perspectives and Current Directions in Hockey Analytics. *Annual Review Statistics and Its Application*. 6:19-36. https://doi.org/10.1146/annurev-statistics-030718-105202
2. John Matizs. 2024. 'Open People's Eyes': How the NHL's Evolved in the Decade of Data. *TheScore*. https://www.thescore.com/nhl/news/2848474/open-peoples-eyes-how-the-nh-ls-evolved-in-the-decade-of-data
3. *MoneyPuck*. https://moneypuck.com/data.htm
4. Matt Larkin. 2026. Top 50 NHL UFAs of 2026: Carlson's Value Skyrockets with Raddysh Off the Board. *Daily Faceoff*. https://www.dailyfaceoff.com/news/top-50-nhl-ufas-2026-rankings-list-carlson-raddysh-trouba-tuch-bobrovsky-jenner-kane-lee-andersson-ovechkin