DATA SCIENCE IN SPORTS ANALYTICS & THE NATIONAL HOCKEY LEAGUE

Ryan T. Glasser
University at Buffalo
December 15, 2023

1. Introduction to Data Science in Sports Analytics

1.1 Field Overview

The field of data science serves to identify trends and patterns amongst data to make decisions that benefit companies, communities of people, or any other entities of interest. One can think of data science as a triangular relationship between three separate disciplines: computer science, probability and statistics, and applied mathematics. The combination of these three domains, along with domain experts, that is those who are familiar with and provide insightful knowledge related to the problem at hand, allows for a relationship that has great implications when it comes to building predictive models and making important and impactful decisions. In order to make informed decisions, the data analysis process is supplemented by programming languages, machine learning algorithms, computational techniques, and statistical models that allow for data scientists and domain experts to not only understand trends and patterns in data, but build predictive models that yield high accuracy and low error rates, perhaps for the use cases of developing models that can detect cancer in tissue via a classification mechanism, predict inventory needs for a company based on customer demand, or in the case of sports analytics and professional hockey, devise models that determine the features that contribute most to a winning hockey team.

When it comes to the data science in the realm of sports analytics, the overall goal is to quantify trends as accurately as possible to make decisions. The nature of sports features the objective outcome of a game, based on the score of that game, and ultimately data scientists and sports teams wish to build models using numerical measures that determine who is still standing at the end of the season, and in the cases of professional hockey, winning the Stanley Cup. There comes a challenge in determining how to operationalize numerous variables that make sense in the context of the problem and have meaning, especially when it comes to categorical data. In any sport, the position of a player, or the outcome of a game, both vital metrics in building predictive models and understanding trends within any given game, are categorical and require careful handling in the preprocessing of data and training of models. In all, there is great value in implementing data science into sports, and with data being ubiquitous, there is plenty of data to go around in understanding the dynamics of the game, as well as player and team performance.

1.2 Evolution of Data in Sports

In terms of the frequency of data that is recorded, it is no secret that the past few decades have featured a deluge of big data, characterized by an extreme increase in the amount of data that is available to data scientists and domain experts in sports for use in analytics. In the early- to mid-20^{th} century, when baseball was the overall dominant sport in North America, and other sports such as football and hockey were just gaining traction, any sort of statistics were documented by humans in attendance and subsequently published in newspapers and other prints, as computers and the internet were not once a thought. As discussed in [1], during the 1940s, documentation of vital statistics in hockey was few and far between, and rarely went beyond keeping track of the metric that determines whether teams won games or not, that is the number of goals scored.

With baseball featuring several batting, pitching, and fielding statistics, there came a time where other sports adopted ways of recording data at different sublevels of the sport. For instance, in hockey, rather than simply recording the number of goals scored in a game or who scored those goals, metrics were devised by focusing on where a goal was scored on the ice, and which players touched or passed the puck before a goal. With such, came additional metrics by the names of slot shots, high danger scoring chances, assists, and points. Consequently, keeping track of more metrics comes more data that is available, and with computing power accelerating over the past few decades and especially within the past few years, there is great possibility in what can be done in collecting data and building appropriate models that can aid in roster and coaching decisions. Beyond all, the quality of data has changed immensely in professional sports. For instance, the National Hockey League (NHL), which once simply recorded the number of goals each team scored to evolving by giving attention to other metrics such as shot locations, takeaways, giveaways, assists, and time

on ice, now showcases its innovation in advanced statistics, coined NHL EDGE. Such a platform features player and puck tracking data, which provides additional aspects of player behavior that had never previously been considered in building models to understand what contributes to team success. For instance, understanding maximum player speed for each team, and which players skate the furthest distance and where on the ice, as seen in [2, Fig. 1] and [2, Fig. 2], respectively, may give insight into which teams succeed, and which do not. Each figure displays data recorded thus far in the 2022-2023 NHL season.

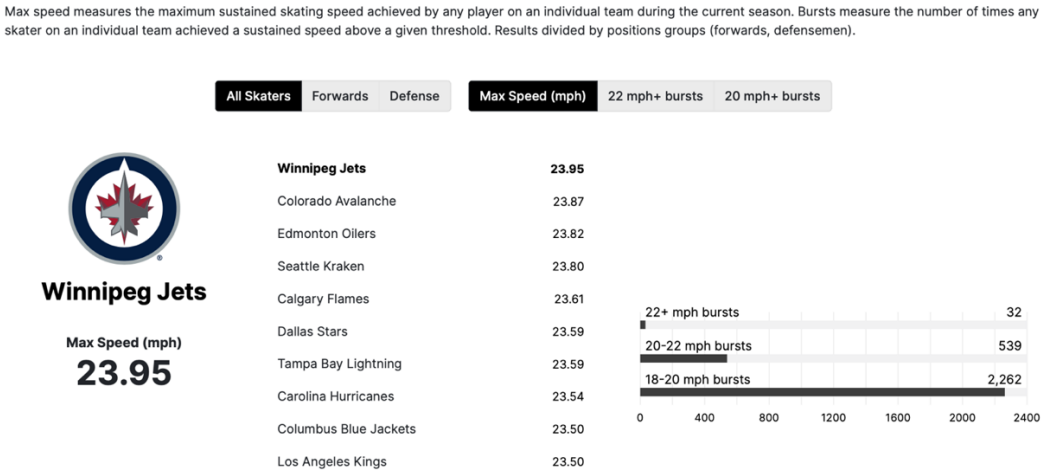

Fig. 1. NHL EDGE data on the 10 fastest NHL teams, determined by the player with the maximum recorded speed on each team.

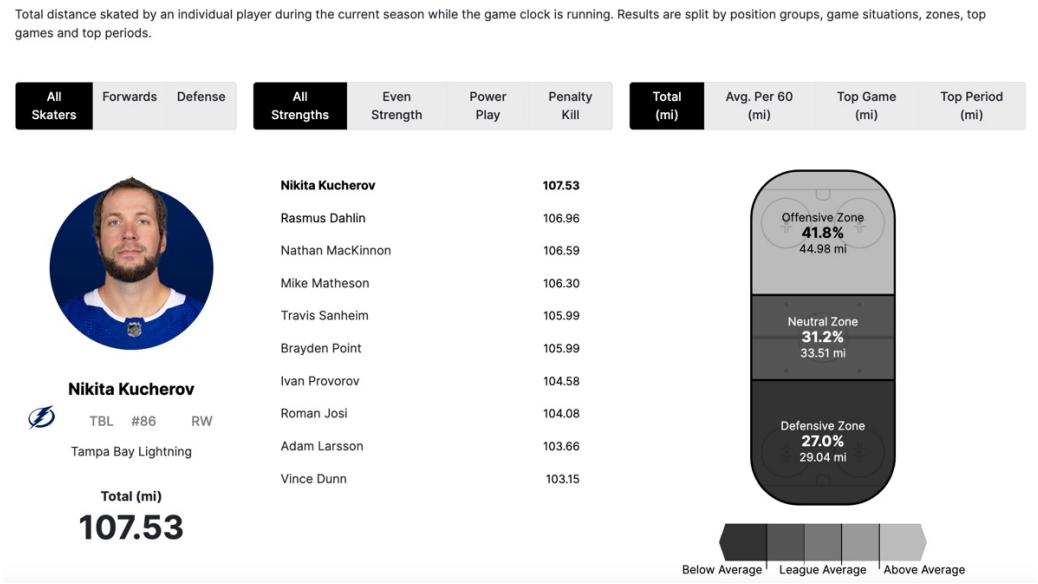

Fig. 2. NHL EDGE data on the 10 NHL players that have skated the farthest, along with a breakdown by zone on the ice.

2. Background

2.1 The NHL

The NHL is a professional ice hockey league in North America, consisting of 32 teams, 25 located in the United States and 7 in Canada. Since its inception in 1917, the NHL has expanded and generated booming revenue, all while attracting new fans to one of the fastest, most physical, and grueling sports. Each team plays a regular season of 82 games consisting of three 20-minute periods, where their goal is to earn points

by winning games and make their way up their respective conference standings to make the playoffs, in which four rounds of best-of-seven matchups determine which team wins the Stanley Cup.

Over the years, the NHL has seen generations of players become more skilled, all while the game becomes faster. With that has come innovative technology and computing mechanisms to improve player and team performance, aid team management in business decisions, such as roster moves, and reach beyond the in-game action by means of fan and community engagement. Managing an NHL franchise is no easy task. From determining lineups on any given night, making trades during the season, scouting prospects, drafting players, and managing team budget and signing players, there are several gears that go into the workings of an NHL team beyond the play on ice that hockey fans witness on a nightly basis. In addition to watching the game and making team decisions based off what is seen, analytics is a viable option that facilitates those decisions as well. With the endless amount of player and game data available today, data scientists in the NHL can leverage the technological advancements of the 21st century to elevate the game and devise strategic plans to build a winning hockey club.

2.2 The Role of a Data Scientist in the NHL

The concept of change in any domain often is met with some sort of resistance, and the same goes with the introduction of analytics into the NHL. The thought of analytics, or the word in general makes many long-standing people in the hockey community cringe, as up until its infusion into the league, the eye test, or the process of making team decisions based on only watching the game in real-time on ice, was the way to go. Analytics was a concept that to many, only required looking at numbers on a spreadsheet, and when a team decision led to more harm than good due to analytical input, resistance only continued. A common misconception is that analytics is the one-size-fits-all method for improving a hockey team or making on- and off-ice decisions. Rather, analytics should be thought of as an additional tool in the toolbox of team management and coaching. Sam Ventura, current Vice President of Hockey Strategy and Research at the Buffalo Sabres, describes the concept of analytics in [3], as a set of information used to better understand the game, and that there are simply some things that cannot be captured through analytics and mathematics and consequently, the eye test is more reliable for. The same goes for the converse, as analytics can provide statistics that one simply watching the game could not reproduce.

When it comes to the connection between the coaching staff and analytics department of an NHL team, it can be described as bidirectional. That is, there is constant communication between both entities before and after each game during the season, and game preparation and performance evaluation are provided to the staff. For instance, upon the completion of a game, a data scientist in the analytics department may deliver the message to the coaching staff that while on the penalty kill, that is while being shorthanded on the ice due to a penalty being committed, the opponent was too easily able to complete cross-ice passes, and the appropriate statistics would be relayed. This in turn allows coaches to make positional adjustments on a game-to-game basis based on the numbers. Beyond that, the collection of such numeric data allows for the derivation of new metrics that can provide more insight and allow data scientists to make predictions on future games within the same season.

Beyond the roles that data scientists in the NHL provide in-season, related to conducting statistical analyses on vital in-game metrics, and relaying those statistics back to other team departments, the responsibility of the offseason largely lies on that of the analytics department. The offseason is a time where teams have the ability to make subtle or drastic changes to their franchise depending on urgency, budget, and overall team performance, and each year, the NHL Entry Draft allows for the work of data scientists to come to fruition. There are a suite of evaluation models and metrics that allow for the precise selection of players for each team based upon the specific needs of a team. The implementation of machine learning algorithms and utilizing statistical metrics to build predictive models that allow teams to arrive at a decision of which players to draft or trade for is essential during this time. Such algorithms include, but are not limited to multiple linear regression, decision trees, logistic regression, K-nearest neighbors, neural networks, and ensemble methods. The draft often is the most demanding time of year for data scientists, due to the several

moving parts leading up to the event, and on the days of the draft themselves. The pool of players that can be selected come from different leagues and countries across the world, as hockey is an international sport, and each has their own style of play. It is inherently difficult to find an objective evaluation of every player that is playing in every league around the world, because players are coached in different ways and play in different leagues. Nevertheless, it is up to the data scientist to build models that best assess players and incorporate not only the natural variability, but cultural and environmental variability amongst players. Beyond the scope of the draft, the tuning of models and hyperparameters, by means of including more or removing features is paramount, as team personnel changes year over year. A changing roster means different player metrics over time, including time on ice, distance traveled, shot locations, and goals, which deserves attention in modifying models that predict player performance.

In terms of specific skills that data scientists in the NHL typically need to establish strategic advantages for their team over the rest of the league, knowledge of machine learning frameworks and programming, especially in Python and/or R, are essential in the cleaning and preprocessing of data, and training and evaluation of models. From there, the final steps of the data analytics process involve the communication and relaying of results. Data analysis is great, but only to the extent that it is interpretable and makes sense to those in the pipeline that are the ones making decisions, and in the case of the NHL, those making team decisions include the general manager, coaches, and trainers. Thus, data scientists in sports analytics and the NHL must be cognizant of how interpretable their analysis is and excel in communicating such results.

2.3 Which Metrics Matter Most?

Hockey is one of the most unpredictable sports amongst the major sports due to the random occurrences that occur within a game. The nature of the sport is chaotic, and it is difficult to predict what will happen when players are skating as fast as they are, some up to 30 miles per hour, pucks are being shot at speeds as high as 100 miles per hour, and the game is being played on a slick sheet of ice. Ultimately, the major metric that matters most to hockey clubs and all sports teams is winning. This translates to either scoring more goals or preventing the opponent from scoring, known as offense or defense, respectively. But there is no analytic usage to using goals for and goals against as a single predictor in a predictive model. To understand what contributes to scoring goals and preventing goals against, it is worth understanding the on-ice plays that correlate with such results. For instance, perhaps it is effective to determine the ability to score goals as a factor of individual player performance and overall team performance. It may be that the locations from which shots are taken on the ice for a team, offensive zone time possession, and the time that a player on the ice has a puck on their stick are important indicators in determining the result of a game.

It is also important to understand that while winning is the major focus point for data scientists in their work for NHL teams, there are other metrics that are of importance for data scientists working for NHL teams. Beyond giving recommendations to coaching staff before and after each game with regards to player and team performance, communicating work to team trainers is also vital for the purposes of keeping a team healthy. It is no coincidence that the teams that can stay healthy through the long 82-game season and four round Stanley Cup playoffs have a higher probability of being the last team standing and winning it all. Thus, health and wellness are of great importance to teams as well, and utilizing machine learning and advanced computing technology to build models to predict player injury has usefulness. If a roster is prone to injury as indicated by models built on previous player data, it might be worth team management considering a change at such positions, as risking players that are susceptible to injuries may do more harm than good in future seasons for a team.

3. Methodologies and Applications of NHL Data Science Problems

3.1 Data Collection and Bias

One of the challenges that faces data scientists in the NHL is trying to sift through the deluge of data that comes from NHL games on a nightly basis. Not only must statisticians and data scientists decipher which

metrics are of importance in building predictive models for their own team, perhaps those that contribute to winning a hockey game or those that determine which player out of a prospect pool best suits the style of a team, but also for the opponents that are faced in at least 82 regular season matchups. As mentioned, these are the problems that many data scientists in analyzing sports do not wish to answer, as they would be giving away work that has attempted to put their team ahead of others. But, one concept known as meta-analysis, that is the ability to evaluate the predictions that are made, attempts to provide a broader understanding of the thinking that many analytics departments take, and is mentioned in a 2021 panel discussion about sports analytics at Carnegie Mellon University [1]. Consider a multiple linear regression model that predicts the number of shots on goal that an opponent will have in a game. Upon computing loss and error rates, it is worth computing the effectiveness of features in predicting what the target variable of interest is. R^2 is a metric that quantifies the proportion of variance in the target variable that is accounted for by the predictor variables. From there, data scientists can analyze the results and determine whether any new information can be derived and if that information is effective in making predictions, or if some features should not be considered.

In terms of the collection of data that is available for predictive modeling and statistical analysis, the majority is widely available to individuals who wish to access the data, from data scientists to the regular fan. Box scores are available after every game, detailing player performance metrics, including goals, assists, shots on goal, hits, and time on ice. Roster information including player demographics, and overall team records are also readily available across decades of NHL seasons. Such statistics can be found through official NHL websites and applications, social media, and many public repositories. In terms of more advanced analytics, as mentioned in Section 1.2, NHL EDGE provides newly implemented puck and player tracking data, allowing for data scientists and those with higher-level understanding to dive deeper and make projections based on additional features, including but not limited to, shot speed, shot location, zone time, skating distance, and skating speed. Upon data collection, programming languages such as Python and R can perform data cleaning, preprocessing, and exploratory data analysis to visualize and understand data further.

In any case of data collection in data science and statistics, it is important to mention the types of biases that may arise from making selections of data to include or not include when conducting data analysis, which are discussed in [4]. First, representation bias or sampling bias becomes an issue in the cases of scouting and during the draft process. Representation bias occurs when certain subpopulations are underrepresented in data. For instance, there may be high caliber, NHL level players in a specific town or region, yet over the past decade, those players may have been neglected due to scouts not visiting those places or watching film. In turn, those players are not considered as much in the building of models that may project that best player for a specific team. The predictions may become skewed in one way or another. Such a problem is also connected to measurement bias, which may occur when scouts make erroneous notes on prospects, or when scorekeepers and statisticians make mistakes while recording in-game metrics, leading to inconsistencies in data and predictive models. Additionally, selection bias may occur when data scientists in the NHL include data in models that is not of importance or exclude data that should be considered in analyzing data and making predictions.

If models are only built on a specific subset or players, but predictions are made on additional types of players, the neglection of some players in the training process can lead to biased results. To combat such biases, it is of utmost importance to understand the composition of the data. For representation bias, being aware of where the data comes from and who collected it is vital. To combat measurement bias, validation systems can be put in place in which statisticians can verify their records against other individuals, as is commonplace in many analytics departments across the NHL today. Finally, selection bias can be curbed by means of exploratory data analysis or evaluation metrics that give insight into which features matter in prediction of a target of interest, and which do not. While a data scientist needs to think about what the driving factors for their specific team are and base decisions off that rather than using all information that is available, selection bias should still be addressed as much as possible.

3.2 Predicting Game Outcome Based on Player and Team Performance

To have a better understanding of the specific deployments of data science in the NHL, the final two subsections describe a few case studies in which machine learning is used to make predictions and aid team management in decisions, the first of which deals with predicting the outcomes of NHL games. It is no secret that trying to formulate a winning team with many moving parts on year-to-year basis is difficult, especially in the randomness of the sport of hockey. A 2018 study conducted by Gu et al. [5] attempts to use historical game data, player performance metrics, as well as information about the opponent to devise a winning formula. Data collection and scraping was done from a variety of NHL sites and sports sites to construct a comprehensive data set using C#. As is typical in other domains that utilize data science, to operationalize and quantify variables, ranking procedures are done, and in this case, player performance is quantified. Statistical analyses, including correlation matrices and hypothesis tests are common in assessing the most impactful features on the dependent variable, that is the game outcome. In professional sports, there are tens of features that can contribute to player performance, both on the offensive and defensive side, which is why the dimensionality reduction technique of Principal Components Analysis (PCA) is useful to not only simplify the model but address multicollinearity. In the study, PCA allowed for the reduction from 18 features to just four components for in characterizing player performance. The same process was carried out on separate metrics for the performance of goaltenders and a team in general, and overall, 19 features were utilized in building predictive models. Beyond this, the nonparametric Wilcoxon's rank-sum test was utilized to better select and devise significant variables in prediction of the outcome of NHL games.

A support vector machine is an effective method to use in binary classification, especially in predicting the outcome of hockey games. A typical setup is $Z = w_1*x_1 + w_2*x_2 + \cdots + w_n*x_n + b$, where a prediction z is a linear combination of the 19 features x_i, multiplied by weights w_i, plus a bias term. Upon training and the arrival at optimal weights, predictions can be made. For the case of the study, if $Z > 0$, the team is projected to win the game, while if $Z < 0$, the team is projected to lose the game. The threshold for classification may vary based on the problem at hand. Upon training on 1,230 regular season games, a test set of the same size yielded 94.05% accuracy [5].

Additionally, ensemble methods are typical when trying to find the best function to represent the target variable by combining several different models, such as naïve Bayes, neural networks, decision tree, and K-nearest neighbors. There are several types of ensemble methods, including Bagging, Boosting, AdaBoost, and RobustBoost, all of which were utilized in trying to increase accuracy when predicting NHL game outcomes. Bagging can aid unstable algorithms by means of more efficient training and predicting, while boosting is an iterative approach that attempts to reduce misclassifications over time [5]. In the case of the Gu et al. study, all ensemble methods outperformed the previously mentioned support vector machine, supporting the idea that ensemble methods can be very useful in hockey analytics.

3.3 Predicting Future Injuries

There is great value in building models that can quantify the injury risk of NHL players, as this aids team management in understanding the conditions of players and making decisions for the betterment of the team. In a 2020 paper by Luu et al. [6] detailing the process of using machine learning to predict injury risk, data collection and scraping of injured players and their corresponding metrics is done via a Python program. Both Python and R helped compile raw data for analyses. Originally collected features were tested for multicollinearity by means of the variance inflation factor (VIF), and as is typical in regression analysis, a variable that yields a VIF factor greater than 10 is excluded.

Machine learning models devised included logistic regression, random forest, K-nearest neighbors, naïve Bayes, XGBoost, and an ensemble modeling for the best three performing singular models. To avoid overfitting the models, as is common in other domains, k-fold cross-validation is employed. In the Luu et al. study, 10-fold cross validation was used on a 90/10 train-test split. Model evaluation metrics included the following: 1) area under the receiver operating characteristic curve (AUC), which is a measure of model

performance, 2) F1 score, a weighted measure of precision and recall, and 3) Brier score loss (BSL), a mean squared error computation. The higher the AUC and F1 scores, and the lower the BSL scores, the better a model performs. In using the 309 observations of injuries for goaltenders and 6,673 injuries for forwards and defensemen to train and evaluate the models, it was found that XGBoost, an algorithm that "uses a gradient boosting framework to solve prediction problems" [6] outperformed all other models, including the commonly used logistic regression algorithm that is employed in many NHL data science related problems. Such results indicate that while regression analysis is simpler and easy to interpret and many may choose to jump at using such an algorithm, it is not always the best solution in addressing data science problems that arise in the realm of hockey analytics, as is indicative in the Luu et al. study that predicts NHL player injuries.

Word Count: 4,264

4. References

[1] "Conversation On Sports + Statistics | By Carnegie Mellon Alumni Association | Facebook," *www.facebook.com*, Jul. 22, 2021. https://www.facebook.com/watch/?v=248682116764328 (accessed Dec. 15, 2023).

[2] EDGE.NHL.com, "NHL EDGE Puck and Player Tracking Statistics - Home," *EDGE.NHL.com*. https://edge.nhl.com/en/home (accessed Dec. 15, 2023).

[3] "Sabres VP of Strategy & Research Sam Ventura On Analytical Approach | The Instigators Overtime Ep. 7," *www.youtube.com*, Dec. 15, 2021. https://www.youtube.com/watch?v=LhzxO3K5oX8 (accessed Dec. 15, 2023).

[4] S. Parvin, "CSE 4/587: Data Intensive Computing Lec 33 : Bias in Data."

[5] W. Gu, K. Foster, J. Shang, and L. Wei, "A game-predicting expert system using big data and machine learning," *Expert Systems with Applications*, vol. 130, pp. 293–305, Sep. 2019, doi: https://doi.org/10.1016/j.eswa.2019.04.025.

[6] B. C. Luu *et al.*, "Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury: An Analysis of 2322 Players From 2007 to 2017," *Orthopaedic Journal of Sports Medicine*, vol. 8, no. 9, p. 232596712095340, Sep. 2020, doi: https://doi.org/10.1177/2325967120953404.