Can Machine Learning Predict Goals in an NHL Season?*

Davis Dunkleberger†

May 10, 2023

Abstract

Analytical decision remains a new trend in sports. Teams will look into new opportunities to push their team to the front lines of trends. Machine learning is a new discipline that can possibly be useful for predicting goals. Scoring goals often leads to more wins. The data is gathered from the hockeyR package. The machine learning algorithms that are used are logit, tree model, neural network, and K nearest neighbor. After running these models, all the models returned the same accuracy of 93%. It shows that machine learning can be useful in hockey decision making. The model and variables used could not be comprehensive enough and lead to under-fitting of the model. More research can be done to create a better fitting model with more variables.

*Thank you to Professor Ransom for helping me nail down this topic of research

†Department of Health and Exercise Science, University of Oklahoma. E-mail address: davisdunk@ou.edu

1 Introduction

Scoring goals is one of the most critical aspects of a successful season to any hockey team. Scoring more goals than any other teams significantly increases the likelihood that team can make the playoffs. Teams will explore every avenue they can to score more goals, win more games, and make the postseason for a chance to win the most prestigious award in hockey: the Stanley Cup.

In the twenty-first century, North American professional sports teams have recognized the emergence of analytics in decision making. Teams are exploring new ways to adjust and gain an advantage over the other teams in their league. Major League Baseball adopted Sabremetrics, the National Football League has adopted a more aggressive fourth down strategy, and the National Basketball Association has started taking more 3s and layups all due to analytical reasoning. The National Hockey League lacks a clear analytical metric like the other three leagues.

With the recent developments of machine learning, a possible hockey application is if it can accurately predict hockey related statistics and outcomes could help front office decision makers make their decisions on roster construction. Establishing an analytical framework for forecasting season-long goal-scoring is a crucial initial step in advancing hockey analytics. Machine learning algorithms can be trained to develop models that accurately identify whether a shot results in a goal or not. Such models could be useful to the front office for long term roster construction. Testing if machine learning models are accurate will help front office decision makers know if it is a viable strategy for roster construction.

2 Literature Review

Analytical decision making has become a popular topic among sports since the early twenty-first century. The book that is credited with popularizing it is *Moneyball* by Lewis (2003). After that book, analytics became a widespread discussion in sports; the premier book on hockey analytics is *Stat Shot* by Vollman, Awad, and Fyffe (2016). Studies on machine learning in hockey have been used to predict tiers of players in research done by Lehmus Persson et al. (2020); similar research has been done by Liu and Schulte (2018) on assigning value to players using neural networks.

Micheal Lewis chronicled the Oakland Athletics' 2002 season in his book *Moneyball*. It is the first seminal work in modern sports analytics. The book followed Billy Beane, general manager of the A's, as he made strategy decisions based on a statistic called on base percentage. The strategy turned many heads within the organization. The idea behind the strategy was that the more a player gets on base, the more runs the team can score, and the more runs a team can score, the more wins that team can accumulate. The A's adopted this strategy out of the economics of baseball; they are a team in a smaller market and have less to spend on payroll than large market teams. Beane found that the players with a high on base percentage were cheaper than the conventional star player at the time. Using this approach, the A's made the playoffs with the lowest payroll in Major League Baseball that year. That offseason, many teams tried to replicate the success of the A's and the sabremetrics era of baseball began.

This story is the first well known analytical story of recent sports teams. This book inspired many people to look at their favorite sport in a different light. Many people, myself included, took a liking to this way of thinking after reading this book. It started the movement on a grassroots level that lead to the current era of sabremetrics in Major League Baseball with launch angle. New

analyst positions were being opened in front offices. Other sports teams began to see the value in the post-*Moneyball* era. The book is important to the topic as it started the trend on analysis on all sports.

While not as detailed as baseball, hockey experienced a book of its own detailing analytics on the rise in the book *Stat Shot* by Vollman, Awad, and Fyffe (2016). *Stat Shot* is a summary book on the statistics that are common among hockey bloggers who cover analytics. These topics include evaluating the value of puck possession, how to best evaluate junior league players on their stats, and lopsided trades by the advanced numbers. The majority of the book, however, is dedicated towards ranking players and evaluating their performance by advanced metrics. This type of analytical research has become popular among the front offices of National Hockey League teams. Roster construction is the first thing any general manager must address on the job. Advanced metrics like the ones in this book should help them make decisions on who to sign and to what value. The metrics in this book are important to the history of hockey analytics. There is no real metric that predicts player or team success. The purpose of testing machine learning is if it can be used to evaluate success or failure.

After the publishing of *Stat Shot*, machine learning took off and research was done to see if it could predict tiers of players in the National Hockey League. This research was done by Lehmus Persson et al. (2020). The authors took statistics for four seasons to see if the five machine learning tribes could predict the ranking of National Hockey League players in the EA NHL game franchise. Using statistics, like assists and goals, and advanced metrics, like Power Play Goals F per 60, the researchers ran through different machine learning models: logistic regression, naive Bayes, Bayesian network, decision tree, K nearest neighbor, and random forest. These models were tasked with classifying forwards, defensemen, and goaltenders into tiers based on different

statistics for each position. The tiers were the top 10%, 25%, and 50%. The results of these models showed that the two Bayesian classifiers were the best predictors of what tier the players are actually in with their actual rankings. This article shows that machine learning can be useful in predicting and classifying players to their EA tier. Hockey analysts could use machine learning to predict hockey related player performance but there is still little research on sport wide questions like the topic of this paper.

The next big machine learning topic used deep learning to create a new Wins Above Replacement metric by Liu and Schulte (2018). The researchers create a wins above replacement metric that they will train a machine using deep learning. Deep learning has been used by hockey analysts to examine the continuous game flow; however, Liu and Schulte focus on the actions of individuals in that moment rather than game state. The deep learning was done using neural networks and placed weight on certain actions over others. These actions include shots, assists, goals, and more. The purpose of this research was to classify players based on their Goal Impact Metric against their salary and see if the players with the most impact are being paid. The research was to evaluate if the best players offensively are being paid as such. Contracts are the best way to compare the market value of players with their impact on the game. The players that have higher game impacts should be paid like it. The researchers also found that their GIM metric was good at predicting which players were on team friendly contracts and would earn a more fitting pay raise on their next deal. The research done by Liu and Schulte is important as it reinforces the idea that machine learning could be used to predict player performance given a set of actions. Machine learning is useful for player evaluation but it is not specifically used to predict if hockey related plays. Machine learning shows it is useful to evaluate players, but there is still more research to be done on it's predictive power.

The research that has been done on machine learning in hockey is focused on player specific evaluations. That is still only a part of the research opportunities for machine learning in hockey. The next big step should be actually predicting on ice events like goals and the focus of this paper is an important part of the future of machine learning in sports.

3 Data

The primary data source for this research is National Hockey League play by play data from the hockeyR package by Morse (2023). The play by play data is publicly available on the League's website. The package scrapes it like an API and loads it into the R software. The data includes 112 columns that denote different aspects of the game state at that time. Some examples of these columns include a description of the event, home and away players on the ice at the time, and how much time is left in the game.

The sheer number of columns of means there are a lot of variables to interpret when it comes to possibly predicting goals. One of the important factors of hockey is special teams, which is known as being on the power play or being short handed. Being "up a man" or "down a man" dramatically shifts how teams take shots. Being on the power play often leads to more shots than normal even strength; inversely, being short handed means you are often defending and could get less shots off. Special teams could influence the shot distance and angle of the shot. Creating a set of dummy variables will help track their significance in if a shot can lead to a goal or not. The strength_code column of the data will denote if it a shot is short handed or on the power play.

The data does not have a column for goals as well. Part of training the machine learning algorithms requires it knows what it needs to look for. Another numeric factor is needed for if a

play is a goal or not. The three factors for power play, short handed, and if a play was a goal are added onto the end of data set. The event_type column will display if a shot was a goal or not.

The next step is to filter down to just plays that result in a shot on goal. This excludes blocked shots as those shots do not result in goals and often go away from the goal once blocked. Filtering to just shots means selecting rows where the event_type column matches either "GOAL", "SHOT", or "MISSED_SHOT". This allows for the data set to only focus on the types of plays that come closest to scoring a goal.

The initial scrape contains all the data from a game and thus includes some text variables. Text variables cannot be used in these models and thus should be removed for simplicity. After filtering down to the shots only plays, removing the unnecessary columns for the regression will put the data set its final form. This step includes removing text variables and unused columns of numeric variables that are not important to the classification models. After removing the unnecessary variables, the data is left with shot distance, shot angle, period, period seconds remaining, short handed, power play, and goal. These variables will be explained in more detail in the empirical methods section.

The last step to create the data set is to drop any columns with NA values. While doing the empirical research, the missing values cannot be estimated and must be dropped. The total number of rows that were dropped with NA values was 300; however, there were still 122,044 observations that were left to observe. That number is plenty to still get a comprehensive look at the trends and create a good predictive model.

4 Empirical Methods

While my approach explores a number of different machine learning algorithms, the primary empirical model can be depicted in the following equation:

$$Goal = \beta_1 SD + \beta 2SA + \beta 3P + \beta 4PSR + \beta 5SH + \beta 6PP + \varepsilon \quad (1)$$

where *Goal* is a binary classification for if a shot is a goal or not, and *SD* is the shot distance from the net, while *SA* is the angle of the shot from the middle of the net when facing it, while *P* is the period the game is in, while *PSR* is period seconds remaining, while *SH* and *PP* are dummy variables that denote if the shot was short handed or on the power play, respectively.

Since goal is a binary classification, the best way to use machine learning to predict if a shot can be a goal is classification models. The model above can be input into all the machine learning algorithms.

The next step is deciding which machine learning algorithms to use to measure accuracy. During the course, three of the five tribes of machine learning stood out to me as possible ones to use. Symbolists, Connectionists, and Analogizers all had algorithms that could predict within certain ranges. Since those three tribes were selected, research would be done using a tree model, a neural network, and a K nearest neighbor regression. A logit model was also used to see if generalized linear model would also be a useful predictor.

While analogizers also use support vector machine in machine learning, it was excluded from this research due to technological restraints. The computer running this analysis is not suited to handle a lengthy analysis like that. The sheer length and complexity of the algorithm will not be possible on the computer that will conduct this research. The other two tribes of machine learning

are hard to calculate on their own and were excluded for that reason.

The first step in starting the analysis is to split the data into training and test sets which will be used to train models and test them, respectively. Then there will be code done to set up each algorithm. This step takes into account the intricacies of the different models that will be run. This includes setting up penalties, setting up regularization parameters, and cross-validations. I chose to do five fold validations as the data set was extremely large.

After each model is set up, the next step is to create a workflow that trains the model on the training set of data. This step also adds the model to find the estimates of the model and remember the relationship between the variables.

After the machine has been trained with that workflow, the code repeats the process but with the tuning parameters that were set up initially. The machine is now testing itself on data after knowing what the training looked like. It is comparing how it did with the training and recording it's accuracy.

As each model is completed, the accuracy that is recorded is stored to put in a table for later. After all the models are completed, the data will be collected into a table for easy reporting of results.

5 Research Findings

The results are reported in Table 1.

The penalties for the three models that use penalties were insignificant as their values had many zeroes before values were displayed. All models did end up with the same accuracy of 0.93. This accuracy equates to 93%. This finding shows that machine learning is an accurate predictor of

goals. The shot distance and shot angle variables were the biggest predictor of if a shot was a goal or not.

All models appeared to return similar accuracy. One of the possible reasons for this is the simplicity of the model itself. There are not a lot of game factors in this model that account for the moving chaos of ice hockey. This lack of variables likely results in under-fitting of the models. There might not be enough variables to evaluate the true accuracy. The simple approach does appear to give credence to the use of machine learning to being useful in hockey analytics.

6 Conclusion

Machine learning showed that it can be a good tool for front office decision makers. The model got it right an exceedingly high percentage of the time. Front office decision makers could evaluate player performance based on plugging a player's shots specifically into this model and see what the machine could predict how many goals that player should have scored. If that player over- or under-performed, it can be used to evaluate what to do with such a player for the next contract. Evaluating players in their usefulness to the organization's goal of winning a Stanley Cup is an important part of the front office's job. Having as many tools as possible to make informed decisions is important going forward.

New analysis tools can assist teams in a big way. Machine learning appears to something National Hockey League teams can help push analytical decision making forward. Machine learning can push player evaluation, a popular topic in hockey analytics, to a new level. Sports often adapt fast to these changes, so being on the cutting edge would be an important and beneficial step for a team to move forward.

References

Lehmus Persson, Timmy, Haris Kozlica, Niklas Carlsson, and Patrick Lambrix. 2020. "Prediction of Tiers in the Ranking of Ice Hockey Players." In *Machine Learning and Data Mining for Sports Analytics: 7th International Workshop, MLSA 2020, Co-located with ECML/PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings*, *Communications in Computer and Information Science*, vol. 1258. Springer, 89. URL `https://doi.org/10.1007/978-3-030-61377-8_8`.

Lewis, Michael. 2003. *Moneyball: The Art of Winning an Unfair Game*. New York: W. W. Norton & Company.

Liu, Guiliang and Oliver Schulte. 2018. "Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation." In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. URL `https://doi.org/10.24963%2Fijcai.2018%2F478`.

Morse, Daniel. 2023. *hockeyR: Collect and Clean Hockey Stats*. URL `https://github.com/danmorse314/hockeyR`. R package version 1.3.1.

Vollman, Rob, Tom Awad, and Iain Fyffe. 2016. *Stat Shot: The Ultimate Guide to Hockey Analytics*. New York: Viking.

Figures and Tables

Table 1: Accuracy of Machine Learning Models

Algorithm	Penalty	Estimate	Cost Complexity	Tree Depth	Min N	Hidden Units
Logit	0.00	0.93	NA	NA	NA	NA
Tree Model	NA	0.93	0.00	10.00	10.00	NA
Neural Net	0.00	0.93	NA	NA	NA	7.00
KNN	0.00	0.93	NA	NA	NA	29.00