A Machine Learning Approach to Predicting Regular Season Success in the National Hockey League

Jacob Morrison **Nafise Faridi Rad**

Abstract

Machine learning is a topic that has recently become prominent in professional sports, and among other applications has been used as a match outcome prediction tool. Despite this movement, the National Hockey League (NHL) has not garnered much attention from the machine learning community. With that in mind a study was completed with the goal of validating machine learning as a match outcome prediction tool for NHL regular season games. The study uses an expanded data set that has never before been employed in this sort of research, along with improved data preprocessing and analysis methods compared to those featured in related works. This led to the confirmation that machine learning is indeed a useful tool in NHL match outcome prediction. Various methods were examined and it was determined that using a Neural Network with 11 hidden layers led to the best prediction, correctly classifying 60% of the 1079 games the model was tested on. This is a very positive result, obtaining a 6% improvement over the baseline level, which arises from making predictions with a decision stump. Further, this study examined which stats are most influential in predicting NHL success and determined that advanced stats such as Corsi and Expected Goals For are most important in match outcome prediction.

1 Introduction

The use of data analytics in the sports world is a relatively new phenomenon, with the origin of the movement believed to have occurred as recently as the year 2000 in Major League Baseball (MLB). Since then every major league professional sports team in North America has implemented an analytics department or has at least employed an analytics expert [9]. As a natural progression of this analytics movement, people involved in major league sports have begun turning to machine learning (ML) as a tool for analyzing the rapidly increasing amount of data available in order to make decisions and predictions, with a particular topic of interest being the prediction of match outcomes [1]. That being said, interest in predicting National Hockey League (NHL) game results though statistical analysis lags behind other professional sports leagues due to the extremely fluid and unpredictable nature of hockey as a sport [3]. This lack of interest is further evidenced by the extremely small number of publications on the topic.

With the ability to accurately predict match outcomes there is the opportunity to realize large financial gain in NHL sports betting, where over 200 million dollars (USD) are expected to be wagered in the next year alone [5]. In addition, knowledge on which types of stats lead to success is crucial in understanding how to design a winning team and lends some insight into the sort of data not currently explored that might aid in future predictions. With that in mind, this study explores the viability of ML in predicting the outcomes of regular season NHL games and attempts to use ML to determine which statistics are most important in predicting success at the NHL level.

Data from all NHL games over the past 3 years was collected, preprocessed, and used to train various ML models for binary classification (Win or Loss). These models were then tested and their accuracy was compared to a baseline prediction formed with a decision stump to validate or

CSPC 532M Final Project Report, Winter Term 1, 2018.

invalidate their effectiveness in NHL match outcome prediction. Finally, feature selection was conducted and the models were retrained and tested to verify the choice of features.

All published attempts to validate machine learning as an NHL match prediction tool to date are unsatisfactory due to a very limited amount of data, both in number of examples and number of features, as well as a lack of hyper parameter tuning. This is further discussed in the next section which reviews related works. Following that the data set used in this study is detailed along with the machine learning methods that were explored. Finally the results are presented and their contributions are discussed along with possible future directions for this research.

2 Related Work

In other professional sports ML has seen success as a tool for predicting match outcomes. A paper by M.C. Purucker was able to predict National Football League (NFL) regular season match outcomes with an accuracy of 78.6% by training a neural network on historical data [8]. Another paper by Loeffelholz et al. was able to predict National Basketball Association (NBA) game results with 74.33% accuracy by again employing a neural network [4]. Further, a 2013 paper by Timmaraju et al. found that training an SVM with an RBF kernel on historical data from the English Premier League (EPL) lead to the best results, achieving 66.7% accuracy at 3 class prediction (Win, Loss, or Tie) [10]. These results indicate that ML is certainly a useful tool when used to predict match outcomes in sports, and that it is worthwhile to examine a few different ML models as there is no obvious best choice.

Now looking specifically at works related to regular season match outcome prediction in the NHL, there have been three papers published to date [7, 11, 12]. The first of these papers was published in 2013 and simply attempted to look at NHL stats and make match outcome predictions. The results indicated a maximum prediction accuracy of 59.38% found with a neural network [12]. The second paper was published in 2014 and attempted to accomplish the same thing, but made predictions based on a combination of machine learning prediction and public sentiment, and were able to achieve an accuracy of 60.25% [11]. The most recent of these papers was also published in 2014 and examined applications in NHL sports betting, though without any improvement in prediction accuracy [7]. Note that all three papers used the same data set. This is because there were no advanced statistics publicly available online until the inception of www.corsica.hockey in 2015. With that in mind, anyone who wanted to use advanced stats had to either collect them game by game throughout an entire season, which is a very time consuming process, or simply use data that someone else had previously gathered.

The data set used by these papers was collected by Weissbock et al. in 2013 solely looking at the 2012-2013 NHL season, which poses a few issues with regards to the validity of the reported results. That season was a lockout shortened season so each team only played 48 games as opposed to a normal season where each team plays 82. This means the data set only had 720 total game examples. A further issue with the examples used is that every single game was considered, including early season games. The problem here is that outliers run rampant in early season games. For example, Imagine team A wins their first game of the season 7-0, and team B loses their first game 3-0. If these two teams then play each other in their second game an algorithm trained only on the stats of their first games will likely predict team A will win. This might be very misleading however, as there is no way team A will win every game of the season 7-0 and maintain a 100% win percentage, just like there is no way team B will lose every game 4-0 and fail to win a single game. With this in mind, early season games should be removed from the data set due to the small sample size that the stats available for those games are based on. Further, because the season in question was lockout shortened, the early season games made up a larger percentage of the total games than normal, making it that much worse that they were included in this data set. Another issue with the data set is that a small amount of features were used. As mentioned, advanced stats were collected manually on a game by game basis meaning every extra stat collected took more effort. With that in mind, only 11 features were used, of which only 3 are considered advanced stats. The difference between regular stats and advanced stats will be explained in the following section.

One last issue that was prevalent across all three of these past papers is that no hyper parameter tuning was involved in the analysis. In all three papers default hyper parameters were employed and models were trained and tested using these. This means the methods were not necessarily optimized and that even with the limited data used the prediction maximum accuracy presented may not represent the actual maximum.

Despite all these issues, one interesting notion presented in the 2014 paper by Weissbock & Inkpen is the idea of a glass ceiling on NHL regular season match outcome prediction accuracy. They argue that due to the randomness of the game of hockey that no model would ever be able to achieve a prediction accuracy above 62% [11]. In this paper this glass ceiling will be examined based on the prediction accuracy of the employed ML methods, while the other issues found in these past papers will be corrected with the goal of providing a more satisfactory result with regards to the validity of machine learning in NHL match outcome prediction.

3 Work Completed

For this study data was collected and used to train and test various ML models. These models were evaluated on their ability to predict the outcomes of NHL regular season games, and feature selection was applied in an attempt to determine which stats are most important for predicting NHL success. In this section the data set used is detailed, the ML methods employed are explained, and the results are presented.

3.1 The Data Set

3.1.1 Explanation of the Data

The data used in this study was taken from all 3720 NHL game over the past 3 full seasons (2015/2016, 2016/2017, and 2017/2018). Data was available dating back further than that, but the game of hockey is one that evolves rapidly over time so the rational is that stats that predict success today are different than stats that predicted success 5 years ago. In table 1 below the 18 features used in this study are outlined. These features include both regular and advanced stats. Advanced stats are simply stats that are not provided by the NHL themselves and that have shown promise in being potentially more useful than traditional stats in evaluating various aspects of a hockey game [6]. Note that a hockey game is typically 60 minutes long, though some go into overtime, so to standardize everything non-percentage stats are shown per 60 minutes of play.

While there are 18 features used, for each game all these features exist for both teams. With that in mind, before preprocessing each example had 36 featues and as mentioned above, there are 3720 examples total. Also note that the features for a given team in a given example are based on the value of that feature for that team over the course of the regular season that example is part of, up to the game representing the example in question. For instance, if Team A plays a game on December 12th, 2015 the Win Percentage used for Team A in that example is their Win Percentage for that season through the end of December 11th.

3.1.2 Data Collection

Data was collected from www.NHL.com and from www.corsica.hockey. From the NHL website data on the winner and loser of each game was collected, along with the general statistics available for each team leading into every game. From the Corsica website advanced statistics for each team prior to every game were obtained. This data collection process was automated with a python script that pulled data from each website at each date of interest to collect all stats up to and including that date. This automation, along with the existence of the Corsica website, made the data collection process much more efficient than the one outlined in the paper by Weissbock et al. that took an entire NHL season [12].

3.1.3 Data Preprocessing

In order to prepare the data to be fed into a ML model it was necessary to complete some preprocessing. First, the features for each example were reduced from 36 to 18 by putting all stats with respect to the home team. This was the method found to work best by Pischedda [7]. What this means is that instead of having two separate features for each stat for the home and away teams, each stat is presented as the home team's value minus the away team's value. For example, if the

Table 1: Features used in the data set

Feature	Name	Description
W%	Win Percentage	Percentage of games played that are won
GDiff	Goal Differential	Difference between goals scored by one team and goals scored against that team
GF%	Goals For Percentage	Percentage of all goals scored in a game featuring a team by that team
PenDIff	Penalty Difference	Number of power plays a team gets minus the number of penelties they take
PP%	Power Play Percentage	Percentage of power plays a teams has that result in a goal for
PK%	Penalty Kill Percentage	Percentage of power plays the other teams has that do not result in a goal against
FOW%	Faceoff Win Percentage	Percentage of all faceoffs in a game won by a team
SF%	Shots For Percentage	Percentage of all shots on goal in a game taken by a team
SH%	Shooting Percentage	Percentage of all shots on goal in a game taken by a team that result in a goal
SV%	Save Percentage	Percentage of all shots on goal in a game taken by the other team that do not result in a goal
CF%	Corsi For Percentage	Percentage of all shots in a game taken by that team
CSH%	Corsi Shooting Percentage	Percentage of all shots in a game featuring a team taken by that team that result in a goal
CSV%	Corsi Save Percentage	Percentage of all shots in a game featuring a team taken by the other team that do not result in a goal
FF%	Fenwick For Percentage	Percentage of all unblocked shots in a game taken by a team
FSH%	Fenwick Shooting Percentage	Perecentage of all unblocked shots in a game taken by a team that result in a goal
FSV%	Fenwick Save Percentage	Percentage of all unblocked shots in a game taken by the other team that do not result in a goal
PDO	Save Percentage Plus Shooting Percentage	This stat serves in part as a measure of luck and is calculated as the sum of a teams save percentage and shooting percentage
xGF%	Expected Goals For Percentage	An advanced stat that expresses the expected goals for percentage for a team based on the quality and quantity of scoring chances both for and against

home team had a win percentage of 76 and the away team had a win percentage of 51 coming into a game, the win percentage feature for that example would be 25.

Secondly, the early season games were removed from the data set. This was done to remove outliers as discussed in the related works section. To be specific, data from the first 21 days of play of each season was removed, corresponding to approximately the first 10 games played by each team. After these two processes were completed the final data set for input into the ML models contained 3236 examples with 18 features each, with each example containing a binary label of either +1 or -1 corresponding to either the home team or the away team winning. An sample example can be seen in table 2 below.

Table 2: A sample example row

Date	Home Team	Away Team	W%	GDiff	GF%	PennDIff	PP%
11/2/2015	Toronto Maple Leafs	Dallas Stars	-0.72	-2.11	-19.4	-15	-6.31
PK%	**FOW%**	**SF%**	**SH%**	**SV%**	**CF%**	**CSH%**	**CSV%**
1.72	0.5	0.81	-5.12	-2.42	1.44	-2.52	-1.61
FF%	**FSH%**	**FSV%**	**PDO%**	**xGF%**	**Result**		
2.98	-3.90	-2.01	-7.53	-7.31	1		

3.2 Machine Learning Methods Implimented

Match outcome prediction can be addressed as a binary classification problem with the output obtained by mapping the home team performance to +1 for winning a game and -1 for losing. The different features collected over the past 3 NHL seasons are then used as input data. In this case both supervised classification and regression methods were employed in an attempt to solve the problem. In this study three specific methods are examined; Neural Networks (NN), Support Vector Machines (SVM) and Random Forests (RF). These methods were chosen as NN and SVM have proven useful in match outcome prediction in other sports [4,8,10], and RF is thought to be the best out of the box classifier [2]. In order to develop these models the order of the examples in the data was randomized and the first third was taken as a training set. The second third was then used as a validation set for hyper parameter tuning and the final third was used as a test set to evaluate the performance of the different models.

Selecting the most relevant features and omitting the irrelevant ones can improve prediction by reducing the test error. Further, it is very useful to know which stats are the most important when it comes to predicting success in the NHL. In this study, the feature selection was done using the Linear Regression (LR) method with L1 regularization. The best model was then retrained with only the selected features and tested in order to confirm that it performed at least as well as it had before implementing feature selection.

3.2.1 Predicting Match Outcomes

The first method used for prediction was RF, with the number of trees being selected through the use of the validation set. A range of trees from 1 to 25 was explored, and using 10 trees resulted in the minimum validation error.

In the second method, SVM, the regularization constant was chosen with validation data as 0.01. The limited domain [0.0001, 10] was used to find this hyper parameter. Next RBF, polynomial and sigmoid kernels were evaluated with the RBF kernel resulting in the lowest validation error.

The final method explored, NN, was trained with the training set for hidden layer values from 1 to 20. Only twenty layers were evaluated to avoid optimization bias and high costs. This hyper parameter was then selected with the validation set and it was determined that using 11 hidden layers minimized the validation error. The eleven hidden layers leads to lower validation error.

Finally, all three models, using the optimal hyper parameters, were evaluated with the test set. The results of the match prediction portion of the study can be seen in table 3 below. Also note that a baseline method was included to compare the results to. The baseline method involved fitting a simple decision stump to the training and data set to give an approximation of the kind of accuracy someone could expect to achieve without implimenting ML methods.

It can be seen in the table that NN led to the best prediction and that all ML methods explored performed better than the baseline on the test set.

Table 3: Match outcome prediction results with various ML methods

Method	Training Error (%)	Validation Error (%)	Test Error (%)
Random Forrest	5	49	45
SVM	45	45	43
Neural Network	38	40	41
Baseline	43	45	46

3.2.2 Feature Selection

The LR method using L1 regularization is a useful tool for feature because it aims to choose only relevant features while minimizing the error. The linear regression with fit intercept was trained and chose 6 features as the most relevant ones. These features were Csh%, Csv%, FSh%, FSv%, SF%, and xGF%. The best model from the previous section, NN, was then retrained using only these features and tested with the test set. Table 4 below shows the results, demonstrating that the feature selection actually helped improve the test error. It can also be seen that removing the irrelevant features increased the training error, which is reasonable when considering fundamental trade off.

Table 4: Match outcome prediction results with NN when applying feature selection

Neural Network	Training Error (%)	Test Error (%)
With Irrelevant Features	38	41
Without Irrelevant Features	39	40

4 Discussion and Future Work

Overall this study was able to achieve 60% accuracy in predicting the results of the 1079 NHL games from the past three NHL seasons that were included in the test set. Further, this accuracy represents a 6% improvement over the baseline level, which represents the prediction accuracy someone would likely be able to achieve without applying ML methods. This is a positive result as it confirms that ML is a useful tool in NHL match outcome prediction that deserves attention from the ML community.

This study also demonstrates, however, that hockey at the NHL level is a very random sport. Despite using a much larger and more complete data set than previous studies, and applying improved analysis methods, this study was unable to improve the match outcome prediction accuracy previously reported. This shows that there will always be some level of unpredictability involved in the NHL. Further, the results of this study strengthen the claim by Weissbock & Inkpen that there is a glass ceiling on the accuracy of NHL match outcome prediction at 62%.

In addition to the match outcome prediction results, the feature selection results proved to be very interesting. Of the 6 features selected, 5 are advanced stats that are not released by the NHL themselves. This defends the use of advanced stats as a performance metric. In addition, 4 of the 6 features selected (CSH%, CSV%, FSH%, FSV%) are highly correlated with puck possession, demonstrating the importance of maintaining puck control for hockey teams. It should also be noted that Corsi stats such as CSH% and CSV%, and Fenwick stats such as Fsh% and FSV% are very similar with the only difference being that one ignores blocked shots and one doesn't. This means it is very hard to say based on the feature selection performed if they are both independently useful stats, or if one is more useful than the other.

While this study does a thorough and satisfying job of confirming the use of machine learning for NHL match outcome prediction, there is always more work that could be done to improve prediction accuracy. With more time it would have been interesting to examine more features that are less widely tracked such as road trip length leading up to the game in question, key roster injuries, if the starting or backup goalie is playing that game, etc. Another interesting idea that could have been implemented with more time is using time period as a hyper parameter. This would mean that when making a prediction for a game only that team's stats over the past k games would be considered, with k being selected by a validation set. This would allow for different trends to come into play and would be more robust to drastic team changes occurring over the course of a season.

References

[1] Alford, H. (2017, May 21). How The Next Generation Of Sports Will Be Driven By Machine Learning. Retrieved from https://blog.starters.co/how-the-next-generation-of-sports-will-be-driven-by-machine-learning-53a0aa3dd812

[2] Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*,15, 3133-3181.

[3] Fraser, A. (2016, September 01). The NHL Is Struggling To Accept Analytics | Articles | Analytics. Retrieved from https://channels.theinnovationenterprise.com/articles/the-nhl-is-struggling-to-accept-analytics

[4] Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA Games Using Neural Networks. *Journal of Quantitative Analysis in Sports,5*(1). doi:10.2202/1559-0410.1156

[5] Novy-Williams, E. (2018, October 29). NHL Embraces Betting in New World Order for Professional Sports. Retrieved from https://www.bloomberg.com/news/articles/2018-10-29/nhl-embraces-betting-in-new-world-order-for-professional-sports

[6] O'Connor, C. (2017, October 09). An advanced stat primer: Understanding basic hockey metrics. Retrieved from https://theathletic.com/121980/2017/10/09/an-advanced-stat-primer-understanding-basic-hockey-metrics/

[7] Pischedda, G. (2014). Predicting NHL Match Outcomes with ML Models. *International Journal of Computer Applications, 101*(9), 15-22. doi:10.5120/17714-8249

[8] Purucker, M. (1996). Neural network quarterbacking. *IEEE Potentials,15*(3), 9-15. doi:10.1109/45.535226

[9] Steinberg, L. (2015, August 18). CHANGING THE GAME: The Rise of Sports Analytics. Retrieved from https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/#25381a1c4c1f

[10] Timmaraju, A.S., Palnitkar, A., Khanna, V. (2013). Game ON! Predicting English Premier League Match Outcomes. *Stanford University*.

[11] Weissbock, J., & Inkpen, D. (2014). Combining Textual Pre-game Reports and Statistical Data for Predicting Success in the National Hockey League. *Canadian Conference on Artificial Intelligence 2014, 8436*, 251-262. doi:https://doi.org/10.1007/978-3-319-06483-3_22

[12] Weissbock, J., Viktor, H., Inkpen, D. (2013). Use of performance metrics to forecast success in the national hockey league. *European Conference on Machine Learning: Sports Analytics and Machine Learning Workshop 2013*