Use of Performance Metrics to Forecast Success in the National Hockey League

Joshua Weissbock, Herna Viktor, and Diana Inkpen

University of Ottawa, Ottawa, Canada
{jweis035, hviktor, Diana.Inkpen}@uottawa.ca

Abstract. Predicting success in hockey is an important area of research which has received little attention in the sports data mining community. We are the first to propose a machine learning approach to forecast success in the National Hockey League. Our approach combines traditional statistics, such as goals for and against, and performance metrics such as possession and luck, in order to build a classification model. We construct several classification models with novel features such as possession and luck in order to build a classification model. Our results indicate that Neural Networks construct the most robust classification models. This confirms the work of earlier researchers, who have also employed Neural Networks in other sports data mining domains. Our results also show the statistics of PDO (which shows, in the short term, the teams playing better or worse than the expected variance) does not aid the prediction.

Keywords: Machine Learning, Hockey, NHL, Classifiers, Neural Networks, Support Vector Machines

1 Introduction

Predicting success in hockey is a subject that has not received much attention compared to other major sports. This may be due to the fact that it is hard to analyze a game of hockey, due to its continuous nature and lack of events (goals). This paper describes a National Hockey League (NHL) Case Study in which we to construct a classification model to predict the outcome of a hockey game. We create classification models to predict success in the National Hockey League (NHL), more specifically, to determine which team is likely to win a game. We use both traditional statistics that are readily available such Goal Differential and Special Teams Success Rate, as well as performance metrics (or "advanced statistics"), used by bloggers and statisticians employed by teams, which have been shown to be more predictive of future success. We further break down these two groups of statistics to see how much they contribute to the success of our classifier.

The rest of this paper is organized as follows. Section 2 provides some background and related research in the field of sports data mining. This is followed, in Section 3, with a discussion of our NHL case study. Section 4 details our experimental approach and results. Section 5 concludes the paper.

2 Background and Related Work

In hockey, five players and a goalie per team are on an ice surface and play for a total of 60 minutes. The goal of the game is to put a rubber puck into the opposing team's net using a 1.5 to 2m long stick made of wood or a composite material. The team who scores the most goals in a game is the winner. In the regular season, if a game is tied after 60 minutes, the teams play an extra 5 minutes of sudden death overtime and after that the game is decided by a shootout. In the playoffs, after 60 minutes, additional 20 minute overtime periods are played until a team scores. As far as the authors are aware, there is no previous work, in the machine learning community, to predict the winner in a hockey game.

Machine learning has been used in other major sports with a varying degree of success to predict the outcome of games, championships and tournaments. Most of the researchers employed neural networks for this task. Chen et al. [1] were among the first to use neural networks for making predictions in sports. They used neural networks to make predictions in greyhound racing and their classifier was shown to be able to make a profit. Huang and Chang [2] used neural networks to make predictions of game winners in the 2006 World Cup and was able to achieve an accuracy of 75%. Purucker [3] used neural networks to make predictions in the National Football League using only four categories he was able to make prediction accuracy of 78.6%. Pardee [4] used neural networks to make predictions for the outcome of the NCAA football bowl games and returned a similar accuracy of 76%. Loeffelholz et al. [5] use neural networks to predict outcomes in the National Basketball Association (NBA) and using common statistics found in the box score of NBA games his model was able to predict with 74.33% accuracy. While neural networks are primarily used in literature, authors have mentioned the use of other classifiers; however, these have not worked as well as neural networks such as [6].

3 National Hockey League Case Study

The NHL is North America's top hockey league comprising of 30 teams: 23 from the United States and 7 from Canada. Teams play a total of 82 games each during the regular season from October to April for a total of 1230 games. There are 2 conferences of 15 teams and each conference is made up of 3 divisions of 5 teams. Within divisions teams play each other 6 times a year, within a conference teams play each other 4 times a year and teams play teams from the other conference 1-2 times a year. At the end of the regular season, the top 8 teams from each conference qualify for the playoffs. The eventual winner wins four best-of-seven series in an elimination tournament and becomes the Stanley Cup Champion.

In our NHL Case Study, data were collected for a period of nearly three months during the 2012-2013 season, for a total of of 517 games between 16 February and 28 April 2013. Due to the lockout this year, this represents about 3/4 of the entire season as teams played 48 games (720 in total). A Python script was created to automate this process, but daily work was required to verify the data and ensure it was collected appropriately. If data were missed, as there is no historical record, it would be difficult to recalculate as it would require iterating through all games and calculating all the statistics.

Goals For	Total number of goals team scored in season (so far).
Goals Against	Total number of goals scored against the team in season (so far).
Goal Differential	Difference between Goals For and Goals Against.
Power Play Success Rate	Ratio where team scored a goal while opposing team had one less man on the ice.
Power Kill Success Rate	Ratio of times team stopped opposing team from scoring while they were down a man due to a penalty.
Shot Percentage	Ratio of goals team scored compared to shots taken.
Save Percentage	Ratio of goals allowed compared to shots stopped by goalie.
Winning Streak	Number of consecutive games won without a loss.
Conference Standings	Team teams current ranking in the standings.
Fenwick Close %	Ratio representing amount of time a team has posession of the puck compared to its opposition.
PDO	Luck, the addition of the teams Sv% and Sh%, over time it regresses to 100%.
5/5 Goals For/Against	Ratio of goals scored by and against team while both teams have 5 players on the ice.

Table 1. All features collected for games.

All statistics collected can be seen in table 1. Recall that some of them were collected before the game, namely traditional and advanced statistics. The traditional statistics are the ones that are easily available from box scores and include Goals For (GF), Goals Against (GA), Goal Differential (GlDiff), Power Play Success Rate (PP%), Power Kill Success Rate (PK%), Shot Percentage (Sh%), Save Percentage (Sv%), Winning Streak and Conference Standing. These were readily available from www.TSN.ca and www.NHL.com. After the game we collected more statistics such as who won and lost, the score, as well as the shots for and against each team. This gives us the power to be able to collect statistics over a smaller subset of games (i.e., the average number of shots against over the last 3 games) instead of seasonal totals and averages. Performance Metrics (or advanced statistics) were also collected before each game. These included Fenwick Close % (a statistic of possession which adds up the total shots, missed shots, and goals for and against); PDO, a statistic of luck; and 5-on-5 Goals For and Against ratio. These statistics are not as readily available and there is no historic record of them. It is only possible to find their current values on websites such as www.behindthenet.ca. Daily work was required to make sure they were collected properly, otherwise the work required to recover their values would be enormous.

Team	Location	Fenwick Close %	Gf	GA	GlDiff	PP%	PK%	Sh%	Sv%	PDO	Win Streak	Conf. Standing	5-5F/A	Class
Toronto	Away	44.92	108	100	8	18.7	85	892	919	1027	2	6	1.05	Win
Ottawa	Home	49.85	89	72	17	29.8	89.4	929	939	1010	3	5	1.12	Loss
Minnesota	Home	48.47	119	126	-7	17.6	80.6	921	911	990	-1	8	0.88	Win
Colorado	Away	46.78	115	149	-34	15.1	80.6	926	909	983	1	15	0.83	Loss
Chicago	Home	55.91	154	99	55	16.9	87	906	928	1022	2	1	1.57	Loss
St. Louis	Away	53.89	126	114	12	19.7	84.5	921	910	989	2	4	1.01	Win

Table 2. Example data for 3 games.

Many of these advanced statistics are just starting to be used by mainstream media and there is evidence that teams are using them to analyze their players; they are also heavily used by bloggers and Internet hockey analysts. They have been shown to be much more predictive of winning, with Fenwick Close having the highest r^2 correlation with points in the standings (0.623 and 0.218 for home and away games) compared to Goals For, Goal Differential, Giveaways, Takeaways, Hits and others [7]. Similarly, looking at the 5-on-5 Goals For/Against ratio, compared to all seasons since 2005, it is founds to have a much higher r^2 correlation with Wins (0.605) and points (0.655) than its traditional statistics counter-parts such as Goals Against / Game, Power Play and Power Kill, and Goals a Game [8].

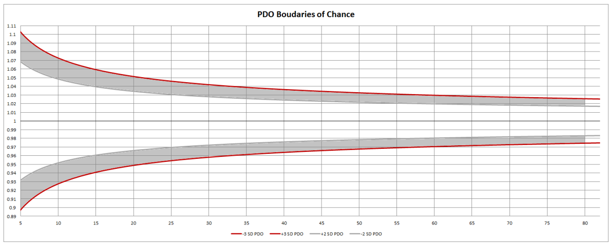

Fig. 1. PDO Boundaries of chance from [9]

Despite its high skill requirement, luck is important to results in the NHL as it makes up 38% of the standings [10]. PDO is an interesting statistic used to analyze hockey. It is not an acronym; rather it is a statistic of luck, luck meaning the results of the gameplay that fall outside of normal boundaries and variance in the players performance [9]. A player cannot maintain a shot percentage that is multiple standard deviations higher or lower than the mean for long periods, nor can a goalie stop every shot that he faces in a season. This is referred to as luck, when the results of the player performance is better (good luck) or worse (bad luck) than the normal average and variance. Hockey seems to be more affected by luck than other major sports due to the low number of events (goals) that happen. A stochastic process can arise that can lead to a goal causing the weaker team to win. Over the long term, luck regresses to the norm, but in the short term you can see which teams have been "luckier" than others. PDO is calculated by the addition of a team's season Sh% and Sv%. Over time, this will regress to 100%; teams who have a PDO higher than 100% have been lucky, while having a PDO less than 100% means that the team has been performing less than its skill level and are seen as unlucky [11]. Over time, teams regress to the norm; within 25 games PDO will be at $100\% \pm 2\%$ [9]. In the short term, we can see who has been lucky.

4 Experimental Design

WEKA [12], a tool that provides many machine learning algorithms, was used for all classification tasks. Preprocessing of the data was done through the entire period of the data collection, as discussed in Section 3. Daily, the new data was collected and it was ensured that the data were valid[1]. Using a python script, the data were represented as the differential between the statistics of the two teams with the winning team receiving the label "Win" and the losing team receiving the label "Loss". As shown in table 2, the first team's data would be in the vector V_1 and the second team's data were in the vector V_2. The Python script calculated $V_1' = V_1 - V_2$ and $V_2' = V_2 - V_1$ and the appropriate Win/Loss labels were attached after[2]. All $517 * 2$ game data vectors were input into WEKA and we used several data mining algorithms. In the first experiment, we looked at how effective traditional, advanced and mixed (both combined) statistics were for predicting success in the NHL. The second part of the experiment further analyzes the "luck" feature to see if can further improve the accuracy.

5 Experimental Results

We consider a binary classification problem in that we want to determine whether a team will win or lose. Using 10-fold cross-validation and a number of WEKA algorithms we input the first three variations of the data sets, namely traditional statistics, advanced statistics and mixed. We used ZeroR to calculate the baseline, this is the results if a classifier were to assign all items to the largest class.

[1] The data sets used in this project are available for future work by others. If you are interested please send a request to the authors by email.

[2] We also modelled the data in a single training example for each game (v.s. one for the win and loss separately) i.e. in the format $V_1 + V_2 + label$ with the label either "HomeWin" or "AwayWin". The results did not vary from our presented results.

For the others we use WEKA's implementation of a neural network (NN) algorithm (good for noisy data); Naive Bayes (NB) (for a probabilistic approach); SMO, WEKA's Support Vector Machine implementation (as it has shown to do well in previous classification tasks); and, J48, WEKA's C4.5 decision tree implementation (as it produces human readable output). All algorithms were tried with their default WEKA parameters in the first experiment.

	Traditional	Advanced	Mixed
Baseline	50.00%	50.00%	50.00%
SMO	58.61%	54.55%	58.61%
NB	57.25%	54.93%	56.77%
J48	55.42%	50.29%	55.51%
NN	57.06%	52.42%	59.38%

Table 3. Accuracies of the first experiment for 10-fold cross-validation.

The best results were achieved when using Neural Networks on the mixed data, as presented in table 3. All algorithms were first tried with their default parameter values from WEKA. We also explored the statistical significance of our results. Our analyses show that, when using the traditional statistics, all algorithms outperformed the baseline. This was not the case when using advanced statistics. When using both statistics, all algorithms (except J48) constructed models that were more accurate the baseline. With additional tuning, using 5 hidden layers and a momentum of 0.11, the NN classifier produced a model with the highest accuracy (59 : 38%); however, there are no statistically difference between the models built by the NN and SVM classifiers. These values were found by inspection. Further tuning of the other classifiers did not result in higher accuracies. The ROC curve can be seen at figure 2. By splitting the data, 66% for training and the remaining 33% for testing we achieve an accuracy of 57.83%. We made sure that no game was split across the two datasets, that is, the two training examples for a game were both in the training or in the test set. We did error analysis and looked at the automatically classified data to see if any pairs have been labeled Win/Win or Loss/Loss. For the Win/Win or Loss/Lass case, we kept the label with the highest confidence the same and inverted the other label, this increased the overall accuracy of the test data to 59% for 66%-33% training/test data. Ensembler learners were also tried using stacking and voting (with 3 and 5 classifiers using the same subset of algorithms) and the accuracy was similar.

When using the Consistency Subset Evaluation (CfsSubsetEval) feature selection method to find which features contribute the most to the model, we are surprised to see the top three are location (being Home or Away), Goals Against and Goal Differential. We are surprised because previous research indicates that performance metrics are more correlated with long term success in the standings [7, 8]. Our further analysis of the results of the classifiers for the first part can be seen in in table 4. Here we can see the precision, recall, f-score and ROC curve for each classifier using 10-fold cross-validation.

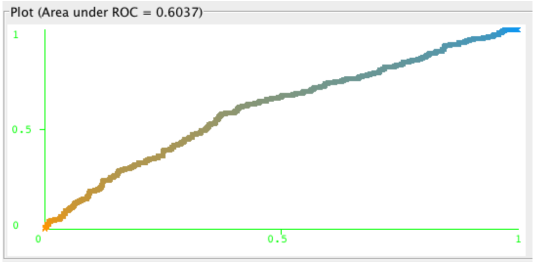

Fig. 2. ROC Curve of tuned Neural Network on the mixed dataset

	Precision	Recall	F-Score	ROC Curve
Baseline	0.5	0.698	0.581	0.5
SMO	0.586	0.586	0.586	0.586
NB	0.567	0.571	0.569	0.588
J48	0.558	0.534	0.545	0.537
NN	0.583	0.594	0.588	0.600

Table 4. Breakdown of each classifier on mixed data for the first experiment using 10-fold cross-validation.

In the second part of our experimental evaluation, we consider the value of PDO in a shortened subset of games. Over the long run, all teams will regress to 100% as players cannot maintain a high (or low) shooting and save percentage for long periods [9]. This is the season value of PDO we used in the first half. We would expect that if we look at the value of PDO over the last n games, we would get a better idea of how lucky a team has been recently, which would be able to help make more accurate predictions. The results of this can be seen in table 5. There does not appear to be any significant change by varying the period of PDO; for the PDOAll and the PDO1 datasets, our statistical significance tests show that all algorithms (except J48) outperform the baseline. This suggests that these performance metrics are not as useful as traditional statistics to predict a single game in the NHL.

5.1 Discussion and Lessons Learned

In the first experiment, we looked at predicting success in the NHL using traditional statistics, performance metrics, and both. After tuning, the best results

	PDO1	PDO3	PDO5	PDO10	PDO25	PDOall
Baseline	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%
SMO	58.61%	58.61%	58.61%	58.61%	58.61%	58.61%
NB	56.38%	56.96%	56.38%	56.58%	56.58%	56.77%
J48	54.93%	55.71%	55.42%	55.90%	55.61%	55.51%
NN	57.64%	56.67%	58.03%	58.03%	57.74%	58.41%

Table 5. Accuracies of the second experiment using 10-fold cross-validation.

come from using Neural Networks with an accuracy of 59.38% (though the difference between NN and SMO was very small). In the long run, it is likely possibly to make a profit off of it. Our results confirm the intuition of earlier researchers to use Neural Networks [1, 2, 13, 3–5]. This choice seems to make the most sense as the algorithm with the best accuracy, as they are known to work well with noisy data such as in sports statistics. What is interesting is that, despite internet hockey analysts showing that performance metrics have a higher correlation with points in the standings, they did not improve our classification rates at all. When we used the Consistency Subset Evaluation feature selection method, it was also interesting to see that the features that added the most value were traditional statistics instead of performance metrics, which have previously shown to be less correlated with points in the standings.

In the second part of our experiments, we considered the predictive power of using a smaller sample size for the PDO statistic of luck. Our results found that using PDO and other performance metrics did not improve our classification results. This seems contrary to hockey analysts who often use statistics such as Fenwick and PDO to predict from mid-season which teams are in playoff standings and will fall (as they tend to have a sub-50% Fenwick with a PDO that is above 100%). The opposite can usually be predicted with teams who have a high Fenwick but a low PDO.

We believe that our model is correct, but we have some future research plans. The first is to aim to increase the accuracy. Other researchers created neural networks to make predictions in other major sports with accuracies in the mid 70s. This may be difficult in hockey, due to luck taking up 38% of the standings. However, we will repeat our experiments in the next hockey season, while collecting additional features and more games, for a large sample size. The additional features that could be tracked that may add value to the classification rate include the number of days of rest between games, time-zone shifts, the affects of long travel, change in altitude, the change in weather, the weather at the arena of the game, gambling odds, injuries on a team, score-adjusted Fenwick, Fenwick when leading or trailing the other team, statistics based on the goal playing, and recent changes in roster are a few that come to mind. Additionally, because of the shortened 2013 season, teams only played 48 games instead of the regular 82. This did not give sufficient time for statistics to regress to the norm and caused surprises in the teams that made it to the playoffs did (such as Toronto with a 44.01% Fenwick-Close), and teams that did not make the playoffs might have (such as New Jersey with a 55.05% Fenwick-Close and a 97.2% PDO). (Note that Toronto was eliminated in the first round of the playoffs.)

6 Conclusion

This paper presented a classifier for the NHL using both traditional, advanced statistics and a mixture. Further examination was given to advanced statistics to see if an improvement in the classifier accuracy could be found.

The best results on our data came from using neural networks with an accuracy of 59.38%. Consistency Subset Evaluation finds that the most important features to the classifier in predicting a single game were location, Goals Against and Goal Differential. While advanced statistics have been shown to make good predictions in the long run (macro scale), traditional stats in this project have performed better in predicting a single game (micro scale).

Future applications of this work would be to aim to predict the winners in the NHL playoffs. Note that the playoffs use four rounds of best-of-seven series, so statistics are more likely to regress to the norm than in a single game. Thus, we are of the opinion that we are more likely to see the better team win. We hypothesize that over a longer series of games you are more likely to see the stochastic processes even out, rather than at the micro scale of a single game, and advanced statistics would show more value. Using a classifier to predict hockey playoff series winner and eventually the Stanley Cup Champion, would be of value to teams as well as to people that bet on games. Another application would be to use betting odds to see if the classifier can make a profit, following the line of thought of Chen et al. [1]. As hockey is somewhat similar to soccer, it would be good to look at machine learning research in soccer for inspiration and see if it would be applicable to hockey.

References

1. Chen, H., Buntin Rinde, P., She, L., Sutjahjo, S., Sommer, C., Neely, D.: Expert prediction, symbolic learning, and neural networks. An experiment on greyhound racing. IEEE Expert **9**(6) (1994) 21–27
2. Huang, K.Y., Chang, W.L.: A neural network method for prediction of 2006 world cup football game. In: Neural Networks (IJCNN), IEEE (2010) 1–8
3. Purucker, M.C.: Neural network quarterbacking. Potentials, IEEE **15**(3) (1996) 9–15
4. Pardee, M.: An artificial neural network approach to college football prediction and ranking. University of Wisconsin (1999)
5. Loeffelholz, B., Bednar, E., Bauer, K.W.: Predicting NBA games using neural networks. Journal of Quantitative Analysis in Sports **5**(1) (2009) 1–15
6. Yang, J.B., Lu, C.H.: Predicting NBA championship by learning from history data. Proceedings of Artificial Intelligence and Machine Learning for Engineering Design (2012)
7. Charron, C.: Breaking news: Puck-possession is important (and nobody told the cbc). http://blogs.thescore.com/nhl/2013/02/25/breaking-news-puck-possession-is-important-and-nobody-told-the-cbc/ (2013) [Online; accessed 12-April-2013].

8. Murphy, B.: Exploring marginal save percentage and if the canucks should trade a goalie. http://www.nucksmisconduct.com/2013/2/13/3987546/exploring-marginal-save-percentage-and-if-the-canucks-should-trade-a (2013) [Online; accessed 12-April-2013].
9. Patrick D: Studying luck & other factors in PDO. http://nhlnumbers.com/2013/1/10/studying-luck-other-factors-in-pdo (2013) [Online; accessed 12-April-2013].
10. Desjardins, G.: Luck in the nhl standings. http://www.arcticicehockey.com/2010/11/22/1826590/luck-in-the-nhl-standings (2010) [Online; accessed 12-April-2013].
11. Charron, C.: PDO explained. http://blogs.thescore.com/nhl/2013/01/21/pdo-explained/ (2013) [Online; accessed 12-April-2013].
12. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
13. Young, W.A., Holland, W.S., Weckman, G.R.: Determining hall of fame status for major league baseball using an artificial neural network. Journal of Quantitative Analysis in Sports **4**(4) (2008) 1–44