Simon Demers*

Riding a probabilistic support vector machine to the Stanley Cup

DOI 10.1515/jqas-2014-0093

Abstract: The predictive performance of various team metrics is compared in the context of 105 best-of-seven national hockey league (NHL) playoff series that took place between 2008 and 2014 inclusively. This analysis provides renewed support for traditional box score statistics such as goal differential, especially in the form of Pythagorean expectations. A parsimonious relevance vector machine (RVM) learning approach is compared with the more common support vector machine (SVM) algorithm. Despite the potential of the RVM approach, the SVM algorithm proved to be superior in the context of hockey playoffs. The probabilistic SVM results are used to derive playoff performance expectations for NHL teams and identify playoff under-achievers and over-achievers. The results suggest that the Arizona Coyotes and the Carolina Hurricanes can both be considered Round 2 over-achievers while the Nashville Predators would be Round 2 under-achievers, even after accounting for several observable team performance metrics and playoff predictors. The Vancouver Canucks came the closest to qualify as Stanley Cup Finals under-achievers after they lost against the Boston Bruins in 2011. Overall, the results tend to support the idea that the NHL fields extremely competitive playoff teams, that chance or other intangible factors play a significant role in NHL playoff outcomes and that playoff upsets will continue to occur regularly.

Keywords: Elo; machine learning; NHL playoffs; relevance vector; support vector.

1 Introduction

The ultimate achievement for a professional sport franchise is to win the league championship after qualifying for the playoffs and surviving successive tournament-style elimination rounds. Although predicting post-season playoff outcomes and identifying factors that increase the likelihood of playoff success are central objectives of sports analytics, few research results have been published on team performance during NHL playoff series. This has left an important knowledge gap, especially in the post-lockout, new-rules era of the NHL. This knowledge gap is especially deplorable because playoff success is pivotal for fans, team members and team owners alike, often both emotionally and economically (Vrooman 2012). A better understanding of playoff success has the potential to deliver new insights for researchers studying sports economics, competitive balance, home ice advantage, home-away playoff series sequencing, clutch performances and player talent.

In the NHL, the Stanley Cup is granted to the winner of the championship after four playoff rounds, each consisting of a best-of-seven series. The NHL uses a home and away sequence where the higher-seeded team based on regular-season points is scheduled to play two games at home and two games away before exchanging home ice advantage for the fifth, sixth and seventh games if required (HH-AA-H-A-H or 2-2-1-1-1 format).

Because the Stanley Cup playoffs are played immediately after the end of a regular 82-game season (unless the season is shortened by a lockout, a player strike or another major event), a wide range of traditional and advanced metrics are available for each team by the time the first playoff round starts. Many of these advanced statistics are reported by mainstream media, analyzed by hockey blog analysts, and used by teams themselves to shape data-driven strategic orientations. Most are now compiled and reported directly by the NHL itself (NHL Insider 2015).

While new playoff prediction models based on these statistics have recently appeared (Emptage 2014b; Luszczyszyn 2014b; Weissbock 2014), sports analytics researchers have generally acknowledged that many analytical tools useful to study and anticipate regular-season team performance do not seem to work as well during the post-season playoffs. In fact, this is a common argument used to try to discredit Moneyball and sports analytics in general.

The analysis that follows is intended to be one more step towards addressing these limitations by evaluating

***Corresponding author: Simon Demers,** Vancouver Police Department – Planning, Research and Audit Section, Vancouver, British Columbia, Canada, Tel.: +604-999-5380, e-mail: simdem@outlook.com

and comparing the predictive performance of various team metrics in the context of NHL playoffs. The analysis strategy consists in compiling various regular-season team metrics and measuring their actual predictive power during playoff series. Combining these features into a probabilistic support vector machine (SVM) and a relevance vector machine (RVM) learning approach, we then derive historical probabilistic playoff performance expectations for NHL teams and identify specific teams that appear to be playoff under-achievers or over-achievers.

2 Theory

For a best-of-seven series, a successful predictor of team strength should be able to identify which team will win four games before the other team. Stronger teams should receive a better rating relative to their opponent and, in turn, teams with a better rating should tend to win their playoff series more often.

Because some randomness remains during a best-of-seven series, upsets are naturally expected and the stronger team will not always prevail as outlined by James, Albert, and Stern (1993). Importantly, this means that even a perfect measure of relative team strength will not be able to explain the outcome of every single playoff series and chance or unexplained variance will still play a role in defining NHL playoff outcomes.

Assuming that each game within a best-of-seven NHL playoff series is played independently of the next one and the favorite team is assigned a constant probability $p \geq 0.5$ of winning any particular game without consideration for the widely assumed home-ice advantage, the probability a of a series upset is given by equation (1).

$$\begin{aligned} a(p) &= 1-p^4\left[1+\binom{4}{1}(1-p)+\binom{5}{2}(1-p)^2+\binom{6}{3}(1-p)^3\right] \\ &= 1-p^4\left[1+\frac{4!}{1!3!}(1-p)+\frac{5!}{2!3!}(1-p)^2+\frac{6!}{3!3!}(1-p)^3\right] \\ &= 1-p^4[1+4(1-p)+10(1-p)^2+20(1-p)^3] \end{aligned} \tag{1}$$

This flows from the fact that there are ${}^{n-1}C_3 = \binom{n-1}{3} = (n-1)!/[3!(n-4)!]$ scenarios under which each team can win the series in n games. Reciprocally, the best team will win the series with a probability of $h(p) = 1 - a(p)$. When $p = 0.60$ for the favorite team in a series, the probability of an upset in that series is approximately $a(0.60) = 1 - 0.7102 = 28.98\%$. Even in a relatively lopsided matchup where $p = 0.75$, the probability of an upset is still approximately 7.06%. When both teams are of equal strength, no predictor would be expected to obtain a success rate above $a(0.50) = 50.00\%$ on average because underdogs would be undifferentiable from favorites.

Two opposite effects are expected to affect the accuracy of playoff series predictions compared to single game predictions. On one hand, actual team performance during best-of-seven playoff series is more likely to regress closer to the true team strength, which means the better team should win best-of-seven series more often than individual games. This is captured by the binomial formula in equation (1). On the other hand, the playoff playing field is more competitive relative to the regular-season. By the time Round 1 starts, most of the weakest teams are already out of the playoffs and only 16 of the strongest NHL teams compete against each other. While weaker teams play against stronger teams during the regular season, there are fewer David vs. Goliath matchups during the playoffs. Everything else being equal, this would tend to reduce our ability to predict playoff outcomes. In addition, home-ice advantage is not as unequivocal or absolute during the playoffs because it alternates between the two opponents (up to Game 7). This tends to muffle somewhat the effects of home-ice advantage, increasing the average series length and adding uncertainty that does not exist for single regular-season games. Finally, various factors such as scheduling and travel during each series are evened out during playoff series. Unlike the regular season, even if back-to-back games, travel distance or jet-lag were factors during Stanley Cup playoffs, they would not offer any significant advantage to any opponent because both teams play the same schedule and travel between the same two arenas during each series. There would only be an effect if one team was systematically able to overcome travel fatigue more effectively. The end result is that playoff predictions are made more difficult by the fact that we miss out on some of the differential information that can be used to predict regular-season games.

How well playoff series can be predicted therefore becomes primarily an empirical question. Using NHL data from 39,563 games that took place during the 1917–2004 period, Ben-Naim, Vazquez, and Redner (2006) found that single game upsets occurred with a frequency of roughly 41.4%. More recently, the best-tuned classification model documented by Weissbock, Viktor, and Inkpen (2013) had an average accuracy of 59.38% for single regular-season games, the *Puck Prediction* model of Emptage (2014a) had an average accuracy of $722/1230 = 58.7\%$ over the 2013–2014 season while the moneyline odds during the same season picked the winning team $728/1230 = 59.2\%$ of the time. If these empirical results reflected the representative probability p that the favorite team will win each individual game of each playoff series, equation (1) predicts that best-of-seven series upsets would occur with a probability $E[\hat{a}]_{Parity\,2006} \approx E[\hat{a}]_{Weissbock\,2013} \approx E[\hat{a}]_{PuckPrediction} \approx E[\hat{a}]_{Moneyline} \approx E[\hat{a}]_{BNVR\,2006} \approx 1-0.70=30\%$. Weissbock (2014) instead suggested that there was a theoretical limit of roughly 62% on single-game predictions. This would translate into a rate of upsets of $E[\hat{a}]_{Theoretical} \approx 1-0.75=25\%$, which means roughly 75% of all playoff series would be *predictable* over the long run.

3 Data

To test known predictors of team performance and establish baseline results, a dataset was compiled with the 105 playoffs series that were played in the NHL between 2008 and 2014 inclusively. The goal in each case is to identify the teams that have better underlying fundamentals because we expect stronger teams to win more often despite any amount of luck that may or may not be involved. A simple classification rule based on these metrics was used to predict which team should be the favorite to win each playoff series. The baseline results are summarized at the end of this section. All the metrics are then incorporated into a machine learning algorithm, which is presented in the following section.

For each best-of-seven series, the following regular-season metrics were assembled for both the winning and the losing team: home-ice advantage as determined by seeding, regular-season shooting percentage ($Sh\% = 100 \cdot GF/SF$ where GF is goals for and SF is shots for on goal); save percentage ($Sv\% = 100(1-GA/SA)$ where GA is goals against and SA is shots against on goal); PDO rating ($PDO = Sh\% + Sv\%$); goal differential ($GD = GF - GA$); goals for percentage ($GF\% = GF/(GF+GA)$); shots for percentage ($SF\% = SF/(SF+SA)$); Corsi percentage ($CF\%$); Fenwick percentage ($FF\%$); score-adjusted Fenwick percentage (SAF); score-adjusted Fenwick percentage in the last 20 regular-season games (SAF_{L20}); total playoff games played; and number of games required to win the previous series. This data was also supplemented by Z-Rating scores, Elo ratings, moneyline betting odds, and EA Sports *NHL* predictions when they were available. All these baseline metrics are well known to and generally well accepted by the sports analytics community. More background information follows on each individual feature.

Home-ice advantage is significant not only because the favorite team will play up to four games at home but because it is granted to the higher-seeded team and therefore it reflects a higher degree of regular-season success. Total playoff games played and playoff games in the previous series are used as proxy measures of player fatigue or rest after Round 1. Previous studies in the context of NBA playoffs have shown team fatigue can affect playoff success in Round 2 even after accounting for relative team strength (Silver 2014) and there are indications it contributes to reinforce home-court advantage (Entine and Small 2008). Anecdotally, it would make sense to expect NHL teams that play more games to also incur more costly injuries as they progress deeper into the playoffs.

Shots differential ($SF\%$) is a traditional hockey metric intended to capture scoring chances by counting shots that are directed towards the net or goalie. Corsi[1] and Fenwick[2] percentages ($CF\%$ and $FF\%$) expand on it by considering *attempted* shots for and against (including shots on goal, missed shots and, specifically for Corsi ratings, blocked shots). Including missed and blocked shots increases the sample size of shot events and arguably provides a more complete measure of on-ice activity. This leads to improved predictive power, especially when data from only a small number of games is available, because random fluctuations in shooting percentage do not have time to even out in the short term (Tulsky 2014b). Attempted shots for have been shown to be strongly correlated with future sustainable goal scoring (Purdy 2010b) and winning percentage (Sullivan 2014), even though they are only somewhat correlated with future playoff prospects (Johnson 2014) and do not imply immediate team success (Charron 2013).

The score-adjusted Fenwick percentage (SAF) builds on the same idea by adjusting for score effects that come up because NHL teams tend to play more aggressively when they trail by at least one goal and more conservatively or defensively when they lead by at least one goal (Costella 2014), especially in the third period (Li 2014). This tends to artificially inflate the rate of attempted shots by teams that trail and play with desperation more often while artificially decreasing the rate of attempted shots by (better) teams that lead more often. Since focusing only on tied score situations reduces the sample size of attempted shots and this can affect the reliability of in-season predictions, Tulsky (2012) introduced a score-adjusted Fenwick (SAF) measure that takes into account all scoring situations. Although score tied and score-adjusted possession

1 The Corsi statistic was named after former NHL goalie and coach Jim Corsi, although blog analyst Tim Barnes is believed to be the first to use it to track offensive dominance.

2 The Fenwick statistic was named after its initial proponent, Battle of Alberta Blog analyst Matt Fenwick.

metrics tend to converge together by the end of the regular NHL season, *SAF* is substantially better at predicting regular-season performance early in the season (when sample size is an issue) and it is currently recognized as one of the best proxy measure of overall team dominance (Yost 2014a).

Despite the predictive value of attempted shots, hockey games are ultimately won or lost based on realized goals. This is why we also include total goal differential (*GD*) and goal percentage (*GF*%) measures. Empirically, goal differential in the NHL has been shown to be closely correlated with current regular-season points (Charron 2011) and somewhat correlated with goal differential in future seasons (Reynolds 2012), although the relationship with playoff success is admittedly less consistent (Emptage 2014e). This was reinforced by the recent findings of Weissbock et al. (2013) in a machine learning context and is supported to a large extent by the excellent track record and lifespan of the Pythagorean expectation theory (James et al. 1993). Interestingly, recent efforts to combine attempted shots and past goals into a composite metric of "weighted shots" (Tango 2014) seem to have resulted in a performance measure that is more closely correlated to future goals than either past goals or attempted shots separately (Cane 2014). This tends to support our approach because it shows that actual goals do add information about team strength, above and beyond attempted shots.

In light of the predictive power of regular-season *GD*, we also include traditional hockey statistics such as regular-season shooting percentage (*Sh*%) and regular-season save percentage (*Sv*%). Both these components are combined into the PDO rating so we consider this metric as well ($PDO = Sh\% + Sv\%$). PDO metrics have been shown to explain *past* performance well but tend to be very poor at predicting *future* performance. Previous analysis work has shown that teams with a higher regular-season shooting percentage (Johnson 2013), higher regular-season save percentage (Desjardins 2010) or higher combined PDO rating (Harding 2015) do not perform significantly better or worse during the playoffs. That being said, the disposition and performance of an NHL goaltender is crucial to his team's success (Swartz et al. 2011), especially during the playoffs (Danco 2013). Similarly, to the extent differences in shooting talent or finishing talent may exist between teams and players, it would make sense to allow for them as part of our analysis. Anecdotally, Luszczyszyn (2014b) obtained outstanding NHL playoff predictions by creating a composite metric that incorporates PDO with traditional special teams percentages and a score-adjusted possession metric.

The score-adjusted Fenwick percentage in the last 20 regular-season games (SAF_{L20}) puts additional weight on recent team dominance, team performance or team momentum effects "down the stretch" that could explain better-than-normal playoff outcomes for a given regular-season performance. The SAF_{L20} rating specifically reflects team performance after the NHL trade deadline and just before the playoffs start, a period that has proven to hold more predictive value than the first half of the season (Desjardins 2011; Tulsky 2014a). As highlighted by Yost (2014b), SAF_{L20} has a good track record of identifying playoff favorites. Roster changes, player injuries and coaching decisions can all affect team performance in the short term, which means a team's performance more than 20–30 games in the past may not reflect adequately its playing ability in the present. The SAF_{L20} ratings are expected to internalize somewhat these eleventh-hour changes and reflect the latest available information on relative team strength. This is why we incorporate them into the analysis.

The Elo ratings were computed using a methodology developed by Bradley and Terry (1952) but adapted for the reality of score-based sports by Hvattum and Arntzen (2010). We implement a goal-based Elo rating system for the NHL with the specific objective of predicting best-of-seven series outcomes during the playoffs. The underlying assumption is that each NHL team possesses a certain level of team strength, which is unknown and fluctuates over time. The objective of the Elo rating system is to estimate team strength by iteratively testing winning expectations and adjusting the rating of each team when their actual performance deviates from these expectations. The Elo rating system therefore tracks each team's playing ability as it changes (Glickman and Jones 1999). The methodology is as follows:

1. At $t = 0$, every Team $i = 1$ to n is assigned an initial rating $R_0^i = 1400$. For analysis purposes, $t = 0$ was defined as October 4, 2005. This is the eve of the 2005–2006 NHL season. As recognized by Hvattum and Arntzen (2010), the Elo ratings cannot be expected to be reliable indicators of team strength until a sufficient number of games have been played. Iterating from the 2005–2006 season allows us to train the goal-based Elo model for three regular seasons before relying on its predictions starting with the 2008 playoffs.
2. The Elo rating of each team is updated after each game t using the adjustment formula represented by equation (2). After game t, the new Elo rating for Team i becomes:

$$R_t^i = R_{t-1}^i + K_t \cdot (W_t^i - p_t^i) \tag{2}$$

3. The attenuation factor K_t reflects the goal difference for game t. The choice of K_t reflects the weight we attribute to each game in terms of how informative it is about relative team strength. Following the goal-based Elo approach proposed by Hvattum and Arntzen (2010), we represent K_t as a dynamic parameter of the form $K_t = K_0(1+\delta_t)^{\lambda}$, where δ_t is the absolute goal difference in game t (unless the game went to overtime, in which case we set $\delta = 0.5$) while $K_0 > 0$ and $\lambda \geq 0$ are fixed parameters that must be chosen. Of note, the traditional Elo system is a special case of this goal-based Elo model with $\lambda = 0$. Some care is required in the choice of K_0 and λ because we want the Elo ratings to adjust themselves as quickly as possible to reflect real fundamental changes in playing ability without introducing unecessary noise through wild fluctuations. After conducting a grid search in the parameter space $K_0 = [2, 30]$ and $\lambda = [0, 2]$, which involved comparing $28 \times 20 = 560$ parameter combinations, we settled on $K_0 = 4$ and $\lambda = 0.2$ since this is the combination that minimized both the least squares loss and the log-loss over the set of individual NHL games that were played during the 2007–2008 to 2013–2014 seasons. This choice of Elo parameters means that $K_t = 4(1+\delta_t)^{0.2}$ in our implementation of the goal-based Elo model detailed by Hvattum and Arntzen (2010).
4. W_t^i reflects the actual outcome of game t. It is equal to 1 if Team i won game t and 0 otherwise.
5. p_t^i is the winning expectancy formula. It reflects the estimated probability that Team i will win game t. We bump the rating of the home team by H to reflect the home-ice advantage that has existed historically. Assuming Team A was the home team and Team B was the visiting team, the winning probabilities are therefore:

$$p_t^A = \frac{1}{1+10^{\frac{R_{t-1}^B - R_{t-1}^A - H}{400}}} \quad \text{and}$$
$$p_t^B = 1 - p_t^A = \frac{1}{1+10^{\frac{R_{t-1}^A - R_{t-1}^B + H}{400}}} \tag{3}$$

6. The home team won 6254 of the 11,339 games in the sample of NHL games downloaded from *Hockey-Reference.com* so we choose H in such a way that when two teams with the same exact rating meet, the home team is expected to win $6254/11{,}339 \approx 55.15\%$ of the time. That is, we choose H such that:

$$p_t^A = \frac{1}{1+10^{-\frac{H}{400}}} = \frac{6254}{11{,}339} \tag{4}$$

Solving for H yields $H = -400 \cdot \log_{10}(11{,}339/6254 - 1) \approx 35.95$ so this is the rating advantage we give to the home team in each individual game matchup.

As noted by Glickman and Jones (1999), the $W_t^i - p_t^i$ term in equation (2) reflects the discrepancy between the expected outcome and the actual outcome. Of note, there is necessarily an adjustment at each iteration because W_t^i can only take a value of 0 or 1 and $0 < p_t^i < 1$.

Although the Elo rating of each team is updated after each game throughout each season (including the playoffs), we are primarily interested to obtain a measure of team strength just before each playoff series starts so these are the ratings we focus on for analysis purposes.

We incorporate moneyline odds into the model as a way to capture public sentiment and the wisdom of the crowd (Emptage 2013). It is a way to capture the opinion of hockey experts (bettors) who have a vested interest in being as accurate as possible (Emptage 2014f). The moneyline betting data reflects the opening series price data for each playoff series. It was compiled from the archives of the *Scores and Odds* website for 2008–2010 (http://www.scoresandodds.com), from the *StanleyCupOdds.ca* and *HockeyOdds.ca* websites for 2011 and Round 1 of the 2012 playoffs, from the *Bleacher Report* website for the subsequent rounds of the 2012 playoffs (http://www.bleacherreport.com) and from the *Sporting News Linemakers* website for 2013–2014 (http://linemakers.sportingnews.com).

After each regular NHL season since at least 2008, the videogame developer EA Sports has published playoff predictions. These predictions are based on a simulation conducted using the *NHL* videogame, one of the flagship game within the EA Sports portfolio. The game simulations rely on state-of-the-art artificial intelligence, real-life data from each active player and realistic line combinations to extrapolate the results of each matchup. In 2008, 2009, 2010, 2013 and 2014, EA Sports published official predictions for the entire playoff bracket before the start of the first round. As a consequence, some of the matchups predicted by EA Sports for the second, third and final rounds did not take place. Exceptionally, in 2011 and 2012, EA Sports published round-by-round predictions. The end result is that EA Sports predictions are available for 79 of the 105 playoff series (including all 56 Round 1 series). For coding purposes, the favorite team according to EA Sports receives a rating of +1 while the EA Sports underdog receives a rating of –1. When no EA Sports prediction is available, both teams receive a neutral (non-informational) rating of 0.

Every metric except the Elo ratings, moneyline odds and some of the EA Sports predictions is based solely on regular-season team performance. Historical playoff results were obtained from the *WhoWins*™ website (http://www.whowins.com). The data on goal differential and regular-season team points was obtained from *NHL.com*. Five-on-five (all scoring situations) *CF*%, *FF*% and SAF ratings are used for analysis purposes. The five-on-five *CF*%, *FF*% and *PDO* ratings were obtained from the *Stats.HockeyAnalysis.com* website. The SAF ratings were obtained from the *Puck Predictions* website (Emptage 2014c). Five-on-five SAF percentages for the last 20 regular season games were assembled from the war-on-ice.com website. The regular-season Z-Ratings for each NHL team were obtained from the *The Z-Ratings Center* website (http://mattcarberry.com/ZRatings/ZRatings.html). The Z-Ratings are based on the same Bradley-Terry system as the Elo ratings but are compiled by an independent source and are included for comparison purposes.

Tables 1 and 2 summarize the success rate of higher-rated teams based on some of the individual predicting variables, first by year and then by playoff round. Because each playoff series has exactly one winner and one loser, the accuracy, precision, recall and specificity of each model are all equivalent and they are henceforth referred to as the percentage correctly classified (PCC) by each predictor.

Only a subset of the most prominent metrics are shown but similar information exists for each feature. In Table 1, the listed years reflect the year when the playoffs took place. For instance, the 2012 results apply to the 2012 NHL playoffs which took place at the end of the 2011–2012 season.

Table 1: Percentage correctly classified (PCC) by playoff year.

	No.	*PDO*	*Elo*	*CF%*	*SAF*	*GD*
2008	15	53.33%	60.00%	53.33%	53.33%	93.33%
2009	15	60.00%	60.00%	60.00%	66.67%	53.33%
2010	15	33.33%	60.00%	73.33%	66.67%	66.67%
2011	15	60.00%	73.33%	53.33%	60.00%	73.33%
2012	15	46.67%	33.33%	53.33%	53.33%	40.00%
2013	15	40.00%	73.33%	66.67%	73.33%	80.00%
2014	15	33.33%	60.00%	73.33%	73.33%	46.67%
Total	105	46.67%	60.00%	61.90%	63.81%	64.76%

Table 2: Percentage correctly classified (PCC) by playoff round 2008–2014.

	No.	*PDO*	*Elo*	*CF%*	*SAF*	*GD*
Round 1	56	48.21%	58.93%	60.71%	66.07%	66.07%
Round 2	28	42.86%	53.57%	64.29%	57.14%	60.71%
Round 3	14	42.86%	78.57%	64.29%	64.29%	71.43%
Round 4	7	57.14%	57.14%	57.14%	71.43%	57.14%
Total	105	46.67%	60.00%	61.90%	63.81%	64.76%

Out of the 105 best-of-seven playoffs series that took place in the NHL between 2008 and 2014 inclusively, teams that had a higher *PDO* rating, a higher *CF*% rating, or a larger regular-season *GD* obtained a success rate of 46.67%, 61.90% and 64.76% respectively against lower-rated opponents. Those teams with a higher *SAF* rating during the regular season won the best-of-seven series 63.81% of the time (67 best-of-seven series out of 105). Teams with the higher Elo rating at the beginning of each series won 60.00% of the time (63 series out of 105). Moneyline favorites, for their part, won 59.62% of their best-of-seven series (62 out of 104 series plus one too-close-to-call series). For baseline comparison purposes, only 57 (54.29%) of the 105 playoff series that were played between 2008 and 2014 were won by the higher-seeded team (who plays on home ice the first two games of the series). This tends to confirm that regular-season standings are poor predictors of playoff performance.

These baseline results do not allow us to conclude that puck possession alone (as measured by attempted shots) can predict the outcome of best-of-seven playoff series consistently better than other regular-season metrics such as *GD*. Although it turned out to be a very poor predictor during the 2012 and 2014 playoffs, *GD* has maintained an honorable track record overall as a playoff series predictor. This is consistent with the results obtained by Weissbock et al. (2013) for single NHL games and it is a testament to the value of the Pythagorean expectation theory by James et al. (1993).

The following section combines all the available metrics into a machine learning algorithm with the objective of testing how well the outcome of NHL playoff series can be anticipated.

4 Methodology

The idea behind the classifier we propose in this section is to create a flexible composite measure of team strength as envisioned by Purdy (2011). Following the machine learning approach pioneered by Weissbock (2014) within the hockey analytics field, we use the predictors introduced in Section 3 to create a new playoff classifier based on a probabilistic support vector machine (SVM) algorithm and a parsimonious relevance vector machine (RVM) learning approach.

The first documented machine learning application for NHL hockey is credited to Weissbock et al. (2013). It was later extended to NHL playoffs by Weissbock (2014) and Computer Scientist (2015). However, to our knowledge, this is the first time RVM is applied in a hockey analytics context. The RVM retains the functional form and predictive performance of the SVM approach but delivers probabilistic predictions, leads to models which are dramatically sparser and avoids the set of free parameters required by the SVM (Tipping 2001). As experienced by Weissbock (2014), the correct choice of kernel parameters is crucial for obtaining good SVM results and this tends to complicate the classification task. The sparsity of the RVM algorithm is especially desirable in the context of NHL playoff data since we are trying to avoid over-fitting a relatively small training set while obtaining generalizable predictions for an equally small validation set. Like SVM, the objective of the RVM algorithm in classification tasks is to find an optimal decision boundary that maximizes the (geometric) margin between winning and losing teams while minimizing a measure of error in the training set. This provides as many correct and confident predictions as possible (meaning on the right side of and far from the separating hyperplane). The RVM algorithm relies on a probabilistic Bayesian learning framework with a Gaussian prior and a flexible linear kernel method identical in functional form to SVM classification. However, the classification formulation also incorporates the Bernoulli distribution and a logistic sigmoid squashing function to deliver consistent RVM estimates of posterior class probabilities (Tipping 2001). The RVM automatically favors simple models that are sufficient to explain the data, marginalizing or pruning out irrelevant variables. Although more complex and more flexible models that are implausibly detailed and over-parameterized might fit the (training) data better, they would interpolate and generalize poorly to the (validation) data. The RVM automatically infers that these unnecessarily complex models are less probable and therefore less desirable (MacKay 1992). Although many of the required computations are analytically intractable and require mathematical approximations, the RVM algorithm produces state-of-the-art results that effectively combine sparsity with predictive accuracy. Fortunately, since our dataset of playoff series is relatively small, we do not have to be concerned with the computational complexity or memory requirements of the RVM algorithm.

Traditional binary classifiers predict a win or a loss given a vector of performance measures. Unlike the probabilistic SVM or RVM approach, they do not provide probabilistic predictions and only provide a hard binary decision. They answer the “who was the best team” and “who will win” questions with complete certainty. This logic is too rigid to cope with the uncertainties of playoff hockey, inherently noisy and inexact hockey analytics data, and non-deterministic nature of the relationship between team strength and playoff outcomes. As demonstrated by MacKay (1992) and Herron (1999), the *moderation* of the classifier’s output has the potential to yield better predictions and allow for more realistic inference. Since there may be regions in the feature space where the classifiers are very uncertain about the applicable class, a deterministic output would likely lead to extreme, unrepresentative and overconfident predictions (MacKay 1992). In the presence of uncertainty, it is preferable to rely on a probabilistic framework to answer the “what was the probability of each team winning” (in-sample) or “what is the probability each team will win” (out-of-sample) questions. Although this is not a new idea (Purdy 2010a), many playoff predictions only come with favorite team picks instead of probabilities. This is the case, for instance, with *EA Sports* predictions, the 2014 NHL.com playoff predictions, the SVM classification results from Weissbock (2014) and most predictions published by hockey blog analysts with some exceptions (Emptage 2014d; Luszczyszyn 2014a; Nilesh 2014).

Observing predictors of team strength (features) and playoff outcomes (class), we want to extract structure from the data in order to conduct some inference (posterior class probabilities) and formulate data-driven predictions (decision value). More specifically, we try to determine how likely it was for the higher-seeded team with the home-ice advantage to win each playoff series. As part of the data preparation, one feature vector was created for each playoff series. This yielded a total of 105 data vectors. In each case, a {0, 1} response variable indicates whether the team with the home-ice advantage won the series. Each vector also contains a total of 33 features. This includes for each of the two opponents in each playoff series: *Sh*%, *Sv*%, *PDO*, *GF*, *GA*, *GD*, *GF*%, *SF*%, *CF*%, *FF*%, *SAF*, SAF_{L20}, the total playoff games played and the total number of games required to win the previous series (when applicable). It also includes the Pythagorean, Elo, Z-Rating and moneyline odds that apply to the higher-seeded team. Finally, when they were available, EA Sports *NHL* predictions were coded as +1 when the higher-seeded team was identifed as the favorite and –1 when the other team was favored.

When possible, estimated series win probabilities are used for RVM and SVM estimation purposes instead of gross ratings. The theoretical Elo win probabilities are calculated using equation (3). These are then converted into approximate theoretical series win probabilities using equation (1). Converting the Z-Ratings into win probabilities is even more straightforward because each team's rating is meant to represent team strength on a multiplicative scale. The odds that Team A will win a game against Team B is given by the ratio $Z_A/(Z_A+Z_B)$, where Z_A is the Z-Rating of Team A and Z_B is the Z-Rating for Team B. This probability can then be converted into approximate theoretical series win probabilities using equation (1).

To convert the moneyline odds data into implicit series win probabilities, the vigorish was first eliminated by implementing the methodology described by Jim Griffin at http://www.onlinebetting.com/remove-vig/. Assuming Team A is the moneyline favorite and its series price is $A<0$ while the series price for Team B is $B>0$, the vigorish percentage is given by $Vig\% = 1/(1+100/|A|)+1/(1+|B|/100)-1$. Of note, this is generally different from taking the midpoint between the posted moneyline odds in absolute value. The implied win probability of the favorite h is then derived by solving the zero-expected profit condition and normalizing to take into account the vigorish, which yields:

$$h=\frac{1}{(1+Vig\%)\cdot(1+100/|A|)} \tag{5}$$

Using the Pythagorean expectation model invented for baseball by Bill James during the early 1980s and recently legitimized in the context of NHL hockey by Dayaratna and Miller (2013), it is also possible to convert regular-season *GD* into a measure of team strength that can be subsequently translated in the form of a win probability for the favorite team in each playoff series. The Pythagorean formula $Pyth_i$ used in equation (6) provides the winning percentage a team should expect to have at a particular point during the season based on goals scored for (*GF*) and goals scored against (*GA*). The twist is that it is used here to estimate the probability of winning the next game (at the margin). In order to convert this measure of team strength into an estimate of win probability p, it is necessary to compare the *relative* Pythagorean expectations of each team involved in each playoff series. This is accomplished by using the *log5* formula attributed to Bill James, which is itself founded on odds ratios. Assuming that Team A is the favorite and Team B is the underdog in a playoff series, the Pythagorean win probability for Team A is assumed to take the form of:

$$p=\frac{Pyth_A(1-Pyth_B)}{Pyth_A(1-Pyth_B)+(1-Pyth_A)Pyth_B} \quad \text{where}$$
$$Pyth_i=\frac{1}{1+\left(\frac{GA+0.5}{GF+0.5}\right)^2} \tag{6}$$

Since modern playoff data is relatively scarce and each playoff year represents one seventh of the data, we kept the 2013 playoff series that followed the shortened 48-game regular season in the data sample for analysis purposes. To ensure comparability with the other seasons, however, we extrapolated the *GA* and *GF* for a 82-game season and recomputed *GD* and Pythagorean odds using these extrapolated values. The idea that a shortened 48-game season would be sufficient to separate true talent from randomness in the NHL has been debated within the hockey analytics field and several authors argued that luck would be likely to supersede talent in a shortened shootout-era season like the 2013 NHL season. As stated earlier, however, a team's performance more than 30 games in the past may not reflect adequately its current playing ability, as demonstrated by the excellent playoff track record of the SAF_{L20} metric (Yost 2014b). Preliminary tests demonstrated that excluding the shortened 2013 season would not affect the predictive power of the composite classifiers in a systematic manner.

We computed probabilistic SVM predictions using the *kernlab* package in R. To prevent overfitting, the SVM classifier is evaluated using a leave-one-out cross-validation procedure. To obtain class probabilities instead of simply generating a binary classification prediction, the SVM estimates were fitted around a sigmoid function (Lin and Weng 2001). The parameters defining the sigmoid function are estimated automatically by minimizing the negative log-likelihood function (Karatzoglou et al. 2004). The *kernlab* package heuristically chooses a sensible sigma value for the Gaussian kernel, which eliminates the need to test several different parameter values.

To create a probabilistic RVM classifier incorporating the known features of each playoff team, we relied on version 2.0 of the *SparseBayes* implementation programmed in MATLAB. This package is an expanded implementation of the algorithm presented in Tipping and Faul (2003). Every feature was rescaled within the range [–1, +1] to avoid numerical difficulties and prevent attributes in smaller numeric ranges from being dominated by those in larger numeric ranges. To prevent overfitting, the RVM classifier is evaluated using the same leave-one-out cross-validation procedure as the SVM classifier. This validation procedure means the RVM algorithm is iteratively trained on 104 playoff series (represented by 104 data vectors) and the resulting model is then validated on the 105th series (represented by the left out data vector). We searched within the relevant range of possible values in order to find a sensible sigma value for the Gaussian kernel of the RVM algorithm and we settled on $\sigma = 4.02$ since it minimized the log-loss over the entire data sample. Testing demonstrated that the RVM predictions would not improve significantly by changing this parameter value.

For comparison purposes, we also estimated classical Probit and Logit regression models. Each of these models included 29 variables instead of the usual 33 features used by the composite machine learning classifiers because $PDO = Sh\% + Sv\%$ and $GD = GF - GA$ would have caused perfect multicollinearity in the regressions. Otherwise, the two regression models are cross-validated using the same leave-one-out approach as the SVM and RVM classifiers.

5 Results

Each classifier delivered cross-validated probabilistic predictions for all 105 playoff series comprising the dataset and we focus on these 105 playoff series to evaluate each classifier.

We compare the SVM and RVM predictions against the Probit and Logit regression results as well as the predictions from four other probabilistic models: moneyline odds, Z-Ratings, Elo ratings and Pythagorean win probabilities.

Let $\hat{h}_i$ be the winning probability assigned to the favorite team in series i by each classifier and let $w_i = 1$ if this team won the series or $w_i = 0$ otherwise. Under the assumption that all playoff series are disputed independently, the likelihood (probability of the data given the model) is given by the product of all series probabilities $\mathscr{L} = \prod \hat{h}_i^{w_i}(1-\hat{h}_i)^{1-w_i}$. It follows that the log-likelihood is given by:

$$\ln \mathscr{L} = \sum w_i \cdot \ln \hat{h}_i + (1-w_i) \cdot \ln(1-\hat{h}_i) \tag{7}$$

Following Glickman (1999), it therefore makes sense to consider a measure of model fit such as:

$$\ell(i)_{LL} = -w_i \cdot \ln \hat{h}_i - (1-w_i) \cdot \ln(1-\hat{h}_i) \tag{8}$$

where $\ell(i)_{LL}$ is the log-loss. This is generally believed to be a natural measure of the goodness of probability estimates because it measures a loss in terms of model log-likelihood.

The logarithm delivers extreme punishments or regret for predictions that are both very confident and wrong. This means risk-averse models that produce fewer really bad over-confident predictions are rewarded, even if it could come at the expense of potentially very good, confident predictions. One major downside is that the $\ell(i)_{LL}$ function can only be computed when the classifier never incorrectly assigns zero probability to the actual outcome (because the penalty is infinite for those cases and the cumulative log-loss becomes undefined). In order to prevent an infinite loss, predictions are capped away from the extremes by a small value. We use the Rule of Three (Eypasch et al. 1995) to cap model predictions for each team within the interval [3/105, 102/105]. This adjustment keeps the log-loss function away from infinity (and absolute zero) in case the predicted probabilities are exactly equal or close to 0% or 100%. Kaggle's 2011 *Photo Quality Prediction*, *Deloitte/FIDE Chess Rating Challenge* and *What Do You Know?* competitions all capped probabilistic predictions within the interval [0.01, 0.99] for the purpose of computing the cumulative log-loss metric. However, our training sample of historical playoff matchups is much smaller and it makes sense to expect a relatively higher floor and lower ceiling given the parity that is believed to exist across modern NHL playoff teams.

In addition to the log-loss function capped to prevent infinite penalties, we also compute two additional measures of fit to compare the goodness of the probability estimates delivered by each classifier: a hinge loss function $\ell(i)_{HL}$ and a least squares loss function $\ell(i)_{LS}$ similar to the modified Huber loss function (Zhang 2004).

The hinge loss $\ell(i)_{HL}$ function is given by:

$$\ell(i)_{HL} = 1 - 4(w_i - 0.5)(\hat{h}_i - 0.5) \tag{9}$$

It follows that the $\ell(i)_{LS}$ loss measure can be calculated using:

$$\ell(i)_{LS} = \left[\frac{\ell(i)_{HL}}{2}\right]^2 = \left[\frac{1-4(w_i-0.5)(\hat{h}_i-0.5)}{2}\right]^2 \tag{10}$$

Across all 105 playoff series, the cumulative loss is given by:

$$\{LL, HL, LS\} = \frac{1}{105}\sum_{i=1}^{105} \ell(i) \tag{11}$$

where $\ell(i)$ is computed using equations (8), (9) or (10), respectively.

The results are reported in Table 3. A smaller cumulative loss or log-loss suggests better model adequacy because the expected distance (Kullback-Leibler divergence) between the predictions and the true distribution is minimized.

The classifiers are listed in decreasing order based on the log-loss. Of note, there is a seemingly monotone relationship between the least squares loss and the log-loss. The Probit and Logit classifiers are by far the ones punished the most harshly by the log-loss function because their over-confident predictions in a few series turned

Table 3: Cumulative loss.

Classifier	*HL*	*LS*	Capped *LL*
Probit	0.8065	0.2992	0.9432
Logit	0.8030	0.2983	0.9291
Z-Rating	0.9149	0.2404	0.6721
Goal-based Elo	0.9301	0.2340	0.6586
RVM	0.9194	0.2303	0.6520
Moneyline	0.9106	0.2265	0.6437
Pythagorean	0.8909	0.2220	0.6351
Probabilistic SVM	0.8659	0.2169	0.6306
Minimum loss	0	0	0.0290
Higher seed ($\hat{h}_i = 0.5429$)	0.9927	0.2482	0.6895
Indecisive ($\hat{h}_i = 0.5$)	1	0.25	0.6931
Maximum loss	2	1	3.5553

out to be wrong. They are the only probabilistic classifiers who were unable to improve overall on a simplistic strategy which involves attributing to the higher seeded team with the home-ice advantage a winning probability of $\hat{h}_i = 0.5429$ to reflect the fact that 57 series out of 105 were won by the higher seeded team during the 2008–2014 period.

Despite the promise offered by its Bayesian approach to machine learning, the RVM had major difficulties dealing with the small unbalanced sample of playoff series and was only able to improve marginally on the Z-Rating and goal-based Elo models. Only the Pythagorean and probabilistic SVM classifiers were able to improve on the moneyline odds.

Overall, the probabilistic SVM classifier proved to be worthy of its reputation and achieved the lowest least squares loss and the lowest log-loss. Although the results are not reported here, the probabilistic SVM classifier also would have achieved the largest percentage correctly predicted (PCP) out of all probabilistic classifiers. It would have achieved an overall success rate of 69.5% (73 properly predicted out of 105 series). The detailed calculations and additional comparisons are available in the accompanying data files.

6 Extensions

Building playoff-ready hockey teams that consistently deliver playoff performances which exceed or are at least aligned with regular-season performance should be a key consideration for NHL team owners, managers and coaches. By comparing the historical playoff performance of each individual NHL team with its expected success rate, it is possible to identify teams that tend to perform significantly better or significantly worse during best-of-seven playoff series for a given level of regular-season performance. Such analysis is the first logical step towards identifying factors that can improve or worsen playoff outcomes for a given level of team quality.

Using the probabilistic prediction results presented in the previous section, it is possible to derive playoff performance expectations for each NHL team and identify specific teams that appear to be playoff under-achievers or playoff over-achievers. Since the probabilistic SVM classifier is demonstrably superior to the other available classifiers, this is the model we use to define historical playoff expectations for each NHL team. Although it is not shown here, similar playoff expectations could be defined using any other probabilistic classifier.

Tables 4–7 list all NHL teams that participated in at least one playoff series during the 2008–2014 period. Table 4 summarizes Round 1 outcomes, Table 5 shows Round 2 outcomes, Table 6 shows Round 3 outcomes and

Table 4: Round 1 playoff performance 2008–2014.

Team	1	2	3	4
	Played	Won	WI 90%	E[SVM]
Anaheim	5	2	[0.71, 3.64]	2.00
Arizona	3	1	[0.23, 2.24]	0.92
Boston	7	5	[2.86, 6.30]	4.34
Buffalo	2	0	[0.00, 1.15]	0.88
Calgary	2	0	[0.00, 1.15]	0.58
Carolina	1	1	[0.27, 1.00]	0.57
Chicago	6	4	[2.08, 5.30]	3.81
Colorado	3	1	[0.23, 2.24]	1.62
Columbus	2	0	[0.00, 1.15]	0.65
Dallas	2	1	[0.24, 1.76]	0.86
Detroit	7	5	[2.86, 6.30]	4.47
Florida	1	0	[0.00, 0.73]	0.27
Los Angeles	5	3	[1.36, 4.29]	2.59
Minnesota	3	1	[0.23, 2.24]	1.02
Montreal	6	3	[1.33, 4.67]	2.52
Nashville	4	2	[0.73, 3.27]	1.38
New Jersey	4	1	[0.23, 2.58]	1.80
NY Islanders	1	0	[0.00, 0.73]	0.31
NY Rangers	6	4	[2.08, 5.30]	3.66
Ottawa	4	1	[0.23, 2.58]	1.50
Philadelphia	6	4	[2.08, 5.30]	2.83
Pittsburgh	7	5	[2.86, 6.30]	4.60
San Jose	7	4	[2.03, 5.69]	4.52
St. Louis	4	1	[0.23, 2.58]	1.40
Tampa Bay	2	1	[0.24, 1.76]	0.93
Toronto	1	0	[0.00, 0.73]	0.39
Vancouver	5	3	[1.36, 4.29]	2.84
Washington	6	3	[1.33, 4.67]	2.75
Overall	112	56		56

Table 5: Round 2 playoff performance 2008–2014.

Team	1 Played	2 Won	3 WI 90%	4 E[SVM]
Anaheim	2	0	[0.00, 1.15]	0.67
Arizona	1	1	[0.27, 1.00]	0.13**
Boston	5	2	[0.71, 3.64]	3.11
Carolina	1	1	[0.27, 1.00]	0.20**
Chicago	4	4	[2.39, 4.00]	2.54
Colorado	1	0	[0.00, 0.73]	0.19
Dallas	1	1	[0.27, 1.00]	0.43
Detroit	5	2	[0.71, 3.64]	3.08
Los Angeles	3	3	[1.58, 3.00]	1.67
Minnesota	1	0	[0.00, 0.73]	0.31
Montreal	3	2	[0.76, 2.77]	1.20
Nashville	2	0	[0.00, 1.15]	1.18*
New Jersey	1	1	[0.27, 1.00]	0.41
NY Rangers	4	2	[0.73, 3.27]	2.44
Ottawa	1	0	[0.00, 0.73]	0.59
Philadelphia	4	2	[0.73, 3.27]	1.78
Pittsburgh	5	3	[1.36, 4.29]	2.38
San Jose	4	2	[0.73, 3.27]	1.83
St. Louis	1	0	[0.00, 0.73]	0.42
Tampa Bay	1	1	[0.27, 1.00]	0.49
Vancouver	3	1	[0.23, 2.24]	1.54
Washington	3	0	[0.00, 1.42]	1.40
Overall	56	28		28

*Significantly different from actual performance at 90% confidence level.
**Significantly different from actual performance at 95% confidence level.

Table 6: Round 3 playoff performance 2008–2014.

Team	1 Played	2 Won	3 WI 90%	4 E[SVM]
Arizona	1	0	[0.00, 0.73]	0.27
Boston	2	2	[0.85, 2.00]	1.13
Carolina	1	0	[0.00, 0.73]	0.47
Chicago	4	2	[0.73, 3.27]	2.42
Dallas	1	0	[0.00, 0.73]	0.22
Detroit	2	2	[0.85, 2.00]	1.14
Los Angeles	3	2	[0.76, 2.77]	1.58
Montreal	2	0	[0.00, 1.15]	0.97
New Jersey	1	1	[0.27, 1.00]	0.42
NY Rangers	2	1	[0.24, 1.76]	1.29
Philadelphia	2	1	[0.24, 1.76]	0.82
Pittsburgh	3	2	[0.76, 2.77]	1.62
San Jose	2	0	[0.00, 1.15]	0.91
Tampa Bay	1	0	[0.00, 0.73]	0.27
Vancouver	1	1	[0.27, 1.00]	0.46
Overall	28	14		14

Table 7 shows Round 4 (Stanley Cup Final) outcomes. The first column of data in each table shows how many times each NHL team reached the applicable round during the 2008–2014 period (seven Stanley Cup championships). The second column shows the realized performance of each NHL team in terms of how many series they actually won within each playoff round. The third column shows the applicable Wilson interval (Wilson 1927) which corresponds to a 90% confidence interval around the number of series won by each team. The fourth column shows how many series wins were expected by the probabilistic SVM classifier.

The Wilson interval (WI) is intended to reflect the relative randomness of playoff success as it relates to true team performance during the playoffs. It is the interval estimator of choice when the sample is small (Vollset 1993; Agresti and Coull 1998; Brown, Cai, and DasGupta 2001) and can be computed using the closed-form formula given by equation (12).

$$\mathrm{WI}(90\%) = \frac{1}{1+1.645^2/n}\left[\bar{h}+\frac{1.645^2}{2n}\pm 1.645\sqrt{\frac{\bar{h}(1-\bar{h})}{n}+\frac{1.645^2}{4n^2}}\right] \tag{12}$$

where n is the total number of playoffs series played by each individual team during the 2008–2014 period, $\bar{h}=\sum h_i/n$ is the percentage of playoff series won by each team and 1.645 is the two-sided 90th percentile Z-score. Although it is not shown, it is straightforward to compute a (looser) WI(95%) using instead the two-sided 95th percentile Z-score (1.960). Equation (12) is expressed in the form of a probability interval so we can multiply it by the number of playoff series each team played in to obtain bounds expressed as a quantity ($n \cdot WI$).

When the number of wins expected by the probabilistic SVM classifier lies outside the Wilson interval, we can conclude that a team's actual playoff performance has deviated from its baseline playoff expectations.

Table 7: Round 4 playoff performance 2008–2014.

Team	1 Played	2 Won	3 WI 90%	4 E[SVM]
Boston	2	1	[0.24, 1.76]	0.66
Chicago	2	2	[0.85, 2.00]	1.41
Detroit	2	1	[0.24, 1.76]	1.25
Los Angeles	2	2	[0.85, 2.00]	1.31
New Jersey	1	0	[0.00, 0.73]	0.25
NY Rangers	1	0	[0.00, 0.73]	0.44
Philadelphia	1	0	[0.00, 0.73]	0.25
Pittsburgh	2	1	[0.24, 1.76]	0.75
Vancouver	1	0	[0.00, 0.73]	0.68
Overall	14	7		7

We conclude the team has under-performed when the expected number of wins is above the top of the Wilson interval or over-performed when the expected number of wins falls below the bottom of the interval.

At the bottom of the pack, the Buffalo Sabres, Calgary Flames, Columbus Blue Jackets, Florida Panthers, New York Islanders and Toronto Maple Leafs did not win any playoff series during the 2008–2014 period. However, their (lackluster) playoff performance was consistent with the (low) expectations of the SVM classifier. As a group, the SVM classifier expected them to win a combined total of approximately three playoff series out of the nine Round 1 series they played in. No NHL team exceeded the Round 1 expectations of the SVM classifier sufficiently to qualify statistically as playoff over-achievers either. Even the Boston Bruins, Detroit Red Wings and Pittsburgh Penguins, who each won 5/7 = 71.4% of their Round 1 series, only met the SVM expectations without exceeding them sufficiently to qualify as Round 1 over-achievers.

The Arizona Coyotes (previously known as the Phoenix Coyotes) and the Carolina Hurricanes both beat terrible SVM odds to win their only Round 2 series and this is sufficient to make them *statistical* playoff over-achievers at the 95% confidence level. Of note, the results tend to hint at the idea that the Chicago Blackhawks and the Los Angeles Kings might also be Round 2 over-achievers at a marginal level of confidence (p-value between 0.10 and 0.15). At the other end of the spectrum, their two series losses qualified the Nashville Predators as Round 2 under-achievers at the 90% confidence interval.

In Round 3 and Round 4, no team was able to separate itself from the SVM expectations sufficiently to qualify as a Stanley Cup over-achiever. That being said, the Vancouver Canucks did come close to qualify as statistical under-achievers after they lost against the Boston Bruins in the 2011 Stanley Cup Final. This is because the SVM classifier attributed them a 68.5% probability of winning the 2011 Finals (p-value between 0.10 and 0.15).

Of note, the New York Rangers outperformed marginally the SVM expectations for Round 1 series but under-performed in all subsequent rounds. On the other hand, the Los Angeles Kings and the Pittsburgh Penguins were the only two NHL teams that outperformed marginally the SVM expectations across all playoff rounds.

7 Conclusion

This paper compares the predictive performance of various team metrics in the context of NHL playoffs. To test known predictors of team performance, a dataset was compiled with the 105 NHL playoffs series that were played between 2008 and 2014 inclusively. The baseline analysis in Tables 1 and 2 provides renewed support for traditional box score statistics, especially in the form of Pythagorean expectations.

Overall, the predictive performance of the probabilistic SVM classifier tends to demonstrate that combining traditional and advanced performance metrics into a new composite metric has the potential to deliver additional predictive power in the context of NHL playoffs. As shown in Table 3, various measures of binomial deviance and model fit tend to favor the SVM classifier. That being said, the results also support the idea that the NHL fields extremely competitive playoff teams, that chance or other intangible factors play a significant role in NHL playoff outcomes, and that playoff upsets will continue to occur regularly.

With a view to identify NHL teams that have proven to be more playoff-ready, the past playoff performance of each NHL team during each individual round was compared with its expected performance based on the SVM classifier. The results in Table 5 show that the Arizona Coyotes and the Carolina Hurricanes can both be considered *statistical* Round 2 over-achievers at the 95% confidence level while the Chicago Blackhawks and the Los Angeles Kings both appear to be Round 2 over-achievers at a marginal level of confidence (p-value between 0.10 and 0.15). For their part, the Nashville Predators would be Round 2 under-achievers at the 90% confidence level. The Vancouver Canucks came the closest to qualify as Stanley Cup Finals under-achievers after they lost against the Boston Bruins in 2011 (p-value between 0.10 and 0.15). Of note, the Los Angeles Kings and the Pittsburgh Penguins were the only two NHL teams that outperformed marginally the SVM expectations across each playoff round over the 2008–2014 period.

As stated by Vrooman (2012), “winning during the regular season is talent driven because good and bad luck usually even out. Winning in the postseason is riskier business because of random elements inherent in short series. […] During the regular season it is certainly better to be good than lucky, but in the playoffs it is probably better to be lucky than good.” Hopefully, this paper shows that the two are not quite as far as many observers think.

References

Agresti, A. and B. A. Coull. 1998. “Approximate Is Better than Exact for Interval Estimation of Binomial Proportions.” *The American Statistician* 52(2):119–126.

Ben-Naim, E., F. Vazquez, and S. Redner. 2006. “Parity and Predictability of Competitions.” *Journal of Quantitative Analysis in Sports* 2(4):1–6.

Bradley, R. A. and M. E. Terry. 1952. “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons.” *Biometrika* 39(3–4):324–345.

Brown, L. D., T. T. Cai, and A. DasGupta. 2001. “Interval Estimation for a Binomial Proportion.” *Statistical Science* 16(2):101–117.

Cane, M. 2014. “Score Adjusted Weighted Shots.” (http://puckplusplus.com/2014/12/10/score-adjusted-weighted-shots/).

Charron, C. 2011. “Pythagorean Expectation and the NHL.” (http://www.nucksmisconduct.com/2011/9/2/2400779/pythagorean-expectation-and-the-nhl).

Charron, C. 2013. “Score Effects and You.” (http://nhlnumbers.com/2013/12/5/score-effects-and-you).

Computer Scientist. 2015. “Can Machine Learning Predict NHL Stanley Cup Champions Using Regular Season Data?” (https://latentbytes.wordpress.com/2015/01/16/can-machine-learning-predict-nhl-stanley-cup-champions-using-regular-season-data/).

Costella, J. L. 2014. “Why Score Effects Are the Key to Your Fancy Stats (Puck Momalytics).” (http://sports.yahoo.com/blogs/puck-daddy/fancy-stats-164244781.html).

Danco, A. 2013. “Predicting the Playoffs Part 2: Does Goaltending Matter?” (http://www.habseyesontheprize.com/2013/6/9/4411742/predicting-the-playoffs-part-2-does-goaltending-matter).

Dayaratna, K. D. and S. J. Miller. 2013. “The Pythagorean Won-Loss Formula and Hockey: A Statistical Justification for Using the Classic Baseball Formula as an Evaluative Tool in Hockey.” *The Hockey Research Journal 2012/13* XVI:193–209.

Desjardins, G. 2010. “Playoff Goaltending and ‘Hot’ Goalies.” (http://www.arcticicehockey.com/2010/4/12/1404784/playoff-goaltending-and-hot-goalies).

Desjardins, G. 2011. “Further to: In-Season Momentum.” (http://www.arcticicehockey.com/2011/3/4/1987571/further-to-in-season-momentum).

Emptage, N. 2013. “The View from Vegas: Crowd-Sourcing Predictions for the NHL Season.” (http://puckprediction.com/2013/09/05/the-view-from-vegas-crowd-sourcing-predictions-for-the-nhl-season/).

Emptage, N. 2014a. “How Did the Puck Prediction Model Do in 2013-14?” (http://puckprediction.com/2014/06/25/how-did-the-puck-prediction-model-do-in-2013-14/).

Emptage, N. 2014b. “Predicting the 2013-14 NHL Playoffs.” (http://puckprediction.com/2014/03/17/predicting-the-2013-14-nhl-playoffs/).

Emptage, N. 2014c. “Score-Adjusted Fenwick: 2007-08 to 2013-14.” (http://puckprediction.com/2014/05/31/score-adjusted-fenwick-2007-08-to-2013-14/).

Emptage, N. 2014d. “The Most Unlikely NHL Playoff Upsets of the Last Five Years.” (http://puckprediction.com/2014/04/04/the-most-unlikely-nhl-playoff-upsets-of-the-last-five-years/).

Emptage, N. 2014e. “What Can Pythagorean Expectation Teach Us About Winning in the NHL?” (http://puckprediction.com/2014/02/24/what-can-pythagorean-expectation-teach-us-about-winning-in-the-nhl/).

Emptage, N. 2014f. “Why Gambling on the NHL is a Terrible Idea.” (http://puckprediction.com/2014/03/15/why-gambling-on-the-nhl-is-a-terrible-idea/).

Entine, O. A. and D. S. Small. 2008. “The Role of Rest in the NBA Home-Court Advantage.” *Journal of Quantitative Analysis in Sports* 4(2):1–9.

Eypasch, E., R. Lefering, C. K. Kum, and H. Troidl. 1995. “Probability of Adverse Events That Have Not Yet Occurred: A Statistical Reminder.” *The BMJ* 311(7005):619–620.

Glickman, M. E. 1999. “Parameter Estimation in Large Dynamic Paired Comparison Experiments.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48(3):377–394.

Glickman, M. E. and A. C. Jones. 1999. “Rating the Chess Rating System.” *Chance* 12(2):21–28.

Harding, J. 2015. “Stats: PDO.” (http://www.beerleagueblog.ca/stats-pdo/).

Herron, M. C. 1999. “Postestimation Uncertainty in Limited Dependent Variable Models.” *Political Analysis* 8(1):83–98.

Hvattum, L. M. and H. Arntzen. 2010. “Using ELO Ratings for Match Result Prediction in Association Football.” *International Journal of Forecasting* 26(3):460–470.

James, B., J. Albert, and H. S. Stern. 1993. “Answering Questions About Baseball Using Statistics.” *Chance* 6(2):17–30.

Johnson, D. 2013. “How Do High Shooting Percentage Teams Perform in Playoffs?” (http://hockeyanalysis.com/2013/04/23/how-do-high-shooting-percentage-teams-perform-in-playoffs/).

Johnson, D. 2014. “Being Honest About Possession Stats as a Predictive Tool.” (http://hockeyanalysis.com/2014/05/11/honest-possession-stats-predictive-tool/).

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis. 2004. “kernlab – An S4 Package for Kernel Methods in R.” *Journal of Statistical Software* 11(9):1–20.

Li, F. 2014. “NHL Play-by-Play Data Mining Part 3C – The Truth About Score Effects.” (https://playfor60minutes.wordpress.com/2014/11/08/110/).

Lin, H.-T. and R. C. Weng. 2001. “A Note on Platt’s Probabilistic Outputs for Support Vector Machines.” (http://www.csie.ntu.edu.tw/~htlin/paper/doc/plattprob.pdf).

Luszczyszyn, D. 2014a. “Playoff Output Projection (POP).” (http://hithisisdom.com/pop/).

Luszczyszyn, D. 2014b. “Predicting the NHL Playoffs.” (http://hithisisdom.com/2014/04/15/predicting-the-nhl-playoffs/).

MacKay, D. J. 1992. “Bayesian Methods for Adaptive Models.” Ph.D. thesis, California Institute of Technology, Pasadena, California.

NHL Insider. 2015. “NHL, SAP Partnership to Lead Statistical Revolution.” (http://www.nhl.com/ice/news.htm?id=754184).

Nilesh. 2014. “FTC’s NHL Playoffs Preview and Predictions.” (http://www.freetankcarter.com/2014/04/ftcs-nhl-playoffs-preview-and.html).

Purdy, T. 2010a. “Playoff Probabilities.” (http://objectivenhl.blogspot.ca/2010/04/playoff-probabilities.html).

Purdy, T. 2010b. “The Relationship Between Outshooting and Outscoring Over Time.” (http://objectivenhl.blogspot.ca/2010/02/relationship-between-outshooting-and.html).

Purdy, T. 2011. “Loose Ends – Part I: Predicting Future Success.” (http://objectivenhl.blogspot.ca/2011/03/loose-ends-part-i-predictive-validity.html).

Reynolds, S. 2012. “Using Goal Differential to Predict the Future.” (http://nhlnumbers.com/2012/11/15/using-goal-differential-to-predict-the-future).

Silver, N. 2014. “Good News: You Won Game 7; Bad News: You’re Less Likely to Win Round 2.” (http://fivethirtyeight.com/features/good-news-you-won-game-7-bad-news-youre-less-likely-to-win-round-2/).

Sullivan, B. 2014. “Projecting 2014-15 NHL Standings Using Fenwick & Corsi Possession Statistics.” (http://www.sportingcharts.com/articles/nhl/projecting-2014-15-nhl-standings-using-fenwick-corsi-possession-statistics.aspx).

Swartz, T. B., A. Tennakoon, F. Nathoo, P. S. Sarohia, and M. Tsao. 2011. “Ups and Downs: Team Performance in Best-of-Seven Playoff Series.” *Journal of Quantitative Analysis in Sports* 7(4):1–17.

Tango, T. M. 2014. “Introducing Weighted Shots Differential (aka Tango).” (http://tangotiger.com/index.php/site/comments/introducing-weighted-shots-differential-aka-tango/).

Tipping, M. E. 2001. “Sparse Bayesian Learning and the Relevance Vector Machine.” *Journal of Machine Learning Research* 1:211–244.

Tipping, M. E. and A. C. Faul. 2003. “Fast Marginal Likelihood Maximisation for Sparse Bayesian Models.” in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume 1, pp. 1–8.

Tulsky, E. 2012. “Score-Adjusted Fenwick Standings Update.” (http://www.broadstreethockey.com/2012/2/22/2816616/score-adjusted-fenwick-standings-update).

Tulsky, E. 2014a. “Quantifying the Added Importance of Recent Data.” (http://www.sbnation.com/nhl/2014/1/21/5329992/nhl-stats-projections-data).

Tulsky, E. 2014b. “Review: Quantifying the Importance of NHL Shot Differential.” (http://www.sbnation.com/nhl/2014/1/28/5353036/nhl-shot-differential-stats-importance).

Vollset, S. E. 1993. “Confidence Intervals for a Binomial Proportion.” *Statistics in Medicine* 12(9):809–824.

Vrooman, J. 2012. “Theory of the Big Dance: The Playoff Payoff in Pro Sports Leagues.” *The Oxford Handbook of Sports Economics*, 1:51–75.

Weissbock, J. 2014. “Forecasting Success in the National Hockey League Using In-Game Statistics and Textual Data.” Ph.D. thesis, University of Ottawa, Ottawa, Canada.

Weissbock, J., H. Viktor, and D. Inkpen. 2013. “Use of Performance Metrics to Forecast Success in the National Hockey League.” in *European Conference on Machine Learning: Sports Analytics and Machine Learning Workshop,* pp. 1–10.

Wilson, E. B. 1927. “Probable Inference, the Law of Succession, and Statistical Inference.” *Journal of the American Statistical Association* 22(158):209–212.

Yost, T. 2014a. “Score-Adjusted Fenwick and Remaining Strength of Schedule.” (http://nhlnumbers.com/2014/1/13/score-adjusted-fenwick-and-remaining-strength-of-schedule).

Yost, T. 2014b. “Stanley Cup Final: Want to Pick Playoff Winners? Try Score-Adjusted Fenwick.” (http://www.sportingnews.com/nhl/story/2014-06-05/stanley-cup-final-rangers-kings-score-adjusted-fenwick-2014-playoffs).

Zhang, T. 2004. “Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms.” in *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, ACM, p. 116.

Supplemental Material: The online version of this article (DOI: 10.1515/jqas-2014-0093) offers supplementary material, available to authorized users.