Expert Systems With Applications 130 (2019) 293–305

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A game-predicting expert system using big data and machine learning

Wei Gu [a], Krista Foster [b], Jennifer Shang [b,*], Lirong Wei [c]

[a] Donlinks School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, P.R. China
[b] Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, United States
[c] Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, United States

ARTICLE INFO

Article history:
Received 14 September 2018
Revised 29 March 2019
Accepted 11 April 2019
Available online 12 April 2019

Keywords:
Expert system
Decision-making
Big data
Machine learning
Ice hockey

ABSTRACT

The National Hockey League (NHL) is a major North American sports organization that earns $3.3 billion in annual revenue, and its stakeholders—team management, advertisers, sports analysts, fans, among others—have vested interest in league competitiveness and team performance. Utilizing player and team data collected from various web sources, we propose an expert system to better predict NHL game outcomes as well as improve recruiting and salary decisions. The system combines principal components analysis, nonparametric statistical analysis, a support vector machine (SVM), and an ensemble machine learning algorithm to predict whether a hockey team will win a game. The ensemble methods improve upon the reference SVM classifier, and the ensemble models' predictive accuracy for the testing set exceeds 90%. The comparison of several ensemble machine learning approaches specifies opportunities to improve the accuracy of game outcome prediction. The system makes it simple for users to employ the learning methodologies and input data sources, evaluate model results, and address the challenges and concerns inherent in predicting hockey game wins.

1. Introduction

The National Hockey League (NHL) is a major North American sports organization, drawing large spectator attendance and commanding an annual revenue of $3.3 billion, including $200 million from TV advertisements. Stakeholders—ownership and management, advertisers, analysts/commentators, fans, among others—have vested interest in league competitiveness and team performance. Extensive stakeholder interest and investment have made sports competitions an important research area, with considerable attention given to understanding why some teams appear to consistently outperform others.

The National Hockey League (NHL) is composed of 30 teams, with 23 located in the United States and 7 in Canada. The league is divided into two conferences, each with three divisions of five teams. Each team plays a total of 82 games in a regular season from October to April, totaling 1230 games for the league. A team faces all other opponents at least once during the regular season, while teams within the same conference play each other four times and those within the same division play each other six times. At the end of the regular season, the top eight teams (six division leaders and two wildcards) qualify for the playoffs. The eventual Stanley Cup® champion wins four best-of-seven series in a single-elimination setting.

Each NHL game consists of three 20 min periods. If a game is tied at the end of regulation, overtime ensues. During the regular seasons from 1999 to 2000 to 2014–2015, overtime play lasted for five minutes with four skaters (plus the goalie) from each side,[1] and the game ended when one team scored in sudden death. Since the 2015–2016 season, the five-minute overtime changed to only three skaters on either side. If there is no scoring during overtime, the game enters a three-round shootout. During each round, one player from each team gets a single shot at the opposing goal. The team with the most goals during the three-round shootout wins the game. If still tied, the teams continue tiebreaker shootout rounds until one team outscores the other. During playoffs, ties result in sudden-death play as opposed to shootouts—multiple 20 min, five-on-five periods are played until one team scores.

Professional sports games call for competent outcome prediction, especially because of the large amount of money involved. Team managers often strive to comprehend and develop strategies

* Corresponding author.
E-mail addresses: guwei@ustb.edu.cn (W. Gu), kmf88@pitt.edu (K. Foster), shang@katz.pitt.edu (J. Shang), weileerong@foxmail.com (L. Wei).

[1] This format was installed in the 2015-2016 season.

https://doi.org/10.1016/j.eswa.2019.04.025
0957-4174/

that are considered necessary in order to win games. Sports games generate enormous amounts of data regarding teams, players, games, and seasons. In recent times, executives are increasingly aware that the immense data could benefit from rigorous study through the application of data science techniques. Management is eager to take advantage of the abundant data by employing machine learning and data mining approaches to help coaches and managers predict game outcomes, assess player performance, evaluate the likelihood of player injury, identify and recruit sports talent, and determine game strategy. In this study, we propose an expert system approach, incorporating big data and machine learning, to predict hockey game outcomes. An expert system can reduce the time it takes to solve the problem and address the concerns more efficiently. Combining machine intelligences and expert knowledge, the proposed system can reduce human error and bias and effectively improve prediction accuracy. Our method offers strategic and competitive advantages that may enhance team management efficiency and increase the likelihood of success.

Developing a winning strategy is a complicated task within team sports, as it involves both intra- and inter-squad player interactions, which are naturally harder to analyze and forecast than the actions of individual competitions (e.g., singles tennis and individual swimming events). Hockey is particularly difficult to predict due to frequent player substitutions, high game speeds, collisions, and fighting. As a result, relative to basketball (NBA) and American football (NFL), the performances of hockey teams and players are more difficult to predict, due to frequent player substitutions, high game speeds, collisions, and fighting.

We scraped data from several websites to construct a database that includes player and team data from NHL regular season and playoff games over the 2007–08 to 2016–17 seasons. There are typically 1230 regular season games, corresponding to 2460 team records. Using an analytical system, we evaluated individual and team performance to identify the factors in winning and develop game outcome prediction. The system combines several methodologies, including principal components analysis (PCA), nonparametric statistical analysis, and support vector machine (SVM) classifier to form an ensemble of machine learning approaches. Given the large dataset, the machine learning and sophisticated data analytics approaches can help uncover hidden patterns, correlations and trends, so as to help management make better decisions. Subsequently, team management can use our decision support tool to further improve performance and maximize wins.

The proposed expert system employs various features: the results of historical matches, player performance indicators, opposition information, etc. The model applies critical analysis with machine learning (ML) and big data analytics to classify and predict hockey game outcomes. The system makes it simple for users to employ the learning methodologies, input data sources, evaluate model results, and address the challenges and concerns that exist in predicting hockey game wins. This ML-based sports prediction framework is novel and serves as a learning strategy. We have found our work to be informative and valuable for research and practice in the sports domain.

In our effort to predict a team's outcome (win/loss), we evaluated the performance of several classifiers: K-nearest neighbor (KNN), support vector machine (SVM), Naïve Bayes, discriminant analysis, and decision trees. We then assessed the effects of ensemble strategies on these classifiers. Namely, we compared boosting, bagging, Adaboost, and RobustBoost. While accuracy increases with the ensemble methods, the preferred strategy depends on the initial classifier. When compared to machine learning models that predict sports outcomes in the existing literature, we have achieved 91.8495% prediction accuracy. This supports our choice of both variables and methods. Further, our approach for predicting a team's outcome demonstrates both the value of merging data from various sources and the benefits of ensemble machine learning methods.

The remainder of the paper is organized as follows: Section 2 briefly introduces the hockey game and NHL rules, then reviews the literature analyzing player performance and prediction techniques in sports; Section 3 discusses the system design; Section 4 details the methodologies used in the construction of our expert system; Section 5 presents the managerial implications and demonstrates how team management's questions can be answered by our system's recommendations; and Section 6 summarizes findings and suggests directions for future research.

The contributions of our research includes: Development of an expert system for sport game outcome prediction; incorporation of machine learning and big data in the expert system and Use of ensemble approach to derive highly accurate results.

2. Literature review

Numerous researchers have analyzed and predicted the outcomes of team sports using machine learning techniques. A sampling of the existing literature is shown in Table 1, indicating the use of neural networks (NN), Naïve Bayes (NB), support vector machines (SVM), decision trees, and discriminant analysis. Although their approaches vary, these researchers all relied on matchup data from previous team encounters and used classifiers to make judgments about which team was better according to teamwork, team performance, and strength. Many only considered a single methodology, and those that compared multiple models did not use ensemble strategies. For example, Huang and Chang (2010) and Huang and Chen (2011) implemented neural networks to predict World Cup football games, and Kahn (2003) applied neural networks to predict American football games, but comparisons to other methods were not discussed. Loeffelholz, Bednar, and Bauer (2009), Miljkovic, Gajic, Kovacevic, and Konjovic (2010), van Roon (2012), and Yang and Lu (2012) applied machine learning methods to predict professional basketball games; while Loeffelholz et al. (2009) used a neural network model, the others utilized Naïve Bayes, a maximum entropy model, and a support vector machine classifier, respectively. Again, these researchers did not compare various approaches. However, Zimmermann, Moorthy, and Shi (2013) compared college basketball game predictions from several machine learning methods, including neural networks (multilayer perceptron), decision trees (C4.5), rule learners (Ripper), Naïve Bayes, and ensemble learners (random forest). They found that the multilayer perceptron and Naïve Bayes models consistently yielded the best predictions.

Despite ice hockey's popularity in Canada and the United States and wealth of game data, the sport has not received as much research attention as basketball (Buttrey, Washburn, & Price, 2011; Carlin, 1996; Kain & Logan, 2014; Rimler, Song, & David, 2010), football (Coleman et al. 2017, Andersson, Edman, & Ekman, 2005; Andersson, Memmert, & Popowicz, 2009), or soccer (Aslan & Inceoglu, 2007; Dijksterhuis, Bos, van der Leij, & van Baaren, 2009). Thus, we aim to make similar predictions for NHL hockey games given the limited existing literature. For instance, Barry and Hartigan (1993), who took goal-scoring to be a Poisson process, proposed a model to estimate the rates at which NHL teams score goals by incorporating factors such as home-ice advantage and manpower (i.e., power play and shots-handed). Morgan, Williams, and Barnes (2013) used hockey to illustrate the use of decision trees and explore attacker-defender interactions. Similarly, Macdonald (2012) stated that faceoffs and hits can predict game scores and noted that many of the predictive variables are correlated. He also gave adjusted, plus-minus estimates based on goals,

Table 1
Comparisons of Methodologies for Predicting Sport Outcomes from the Literature.

Paper Work	Sports	Neural networks	Naive Bayes	SVM	Tree	Discriminant	Ensemble
(Weissbock & Inkpen, 2014)	Ice hockey (NHL)	X	X	X	X		
(Morgan et al., 2013)	Ice hockey				X		
(Huang & Chang, 2010)	Soccer (World Cup)	X					
(Kahn, 2003)	Football	X					
(Loeffelholz et al., 2009)	Basketball	X					
(Miljkovic et al., 2010)	Basketball		X				
(Yang & Lu, 2012)	Basketball			X			
(Weissbock et al., 2013)	Ice hockey (NHL)	X	X	X	X		
Zimmermann et al. (2013)		X	X		X		
Our approach	Ice Hockey (NHL)		**X**	**X**	**X**	**X**	**X**

shots, Fenwick rating,[2] and Corsi rating[3] to measure the value of a player to his team.

Weissbock and Inkpen (2014) used keyword analysis of pre-game reports and in-game statistics to predict game outcomes. Tarter et al. (2009) used NHL performance data to rank hockey players and predict player value for the team. In addition, Feltz and Lirgg (1998) studied collegiate ice hockey for one season. They used efficacy and judgment of players' capabilities to derive correlation between players, team efficacy, and team performance. Leard and Doyle (2011) tested factors of home advantage, momentum, and likelihood of winning on NHL game outcomes.

Other studies focus instead on individual player performances. Voyer and Wright (1998) examined 740 players' performance in scoring, shooting, getting the puck, etc. Perlini and Halverson (2006) evaluated players based on drafting criteria, such as size/strength, skating/speed/power of stride, and shots/scoring. They showed that years-since-draft is the most dominant factor affecting performance, while draft rank is the weakest factor. Emotional intelligence, intrapersonal competency, and general mood were observed to be equally important. However, they did not study the team performance and did not use an analytical approach to synthesize all data into a composite score, as we propose in this research.

Current literature is scant in detailing factors and applying data analytics to predict NHL game outcomes, and to the best of our knowledge, very few researchers have studied the effects of player performance on salaries (Von Allmen, Leeds, & Malakorn, 2015). In this research, we comprehensively examine key factors and their respective relationships with player performance and hockey game victories. We also consider managerial implications, in particular, player compensation and team revenue.

3. System design

3.1. Data collection procedure

Since there is not a single website that has complete player and team information, we combined data from NHL.com, ESPN.com, war-on-ice.com, sn.ca, and hockey-reference.com from the 2007–08 to 2016–17 seasons. As the structure and format of the data vary from site to site, we developed a C# program module to automate the time-consuming task of website crawling and data scraping. Fig. 1 illustrates the data collection process flow.

Fig. 2 shows how our system organizes the head-to-head performance metrics from a game between the Pittsburgh Penguins and Columbus Blue Jackets. Summary tables and pie charts facilitate visualization of the relative team performances for the game.

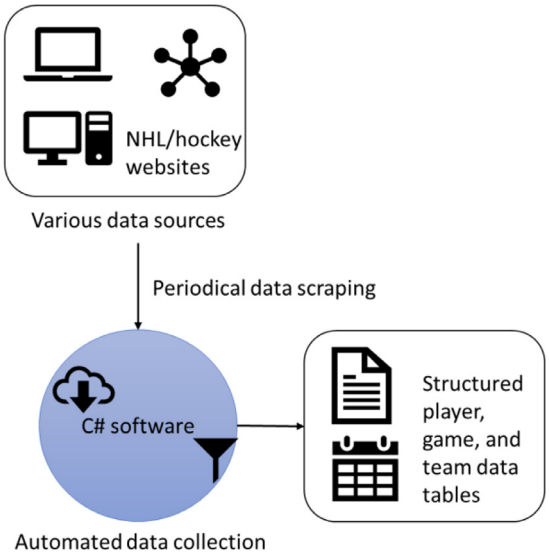

Fig. 1. Flow of Data Collection.

3.2. Analyzing player performance and predicting game winners

To evaluate player performance, we applied a multitude of statistical analyses to rank players and teams. We divided the players into two groups, skaters and goalies, and ranked players for both regular season and post-season play. The sequence is outlined in Fig. 3.

Fig. 4 shows how our system organizes the player performance metrics for all NHL teams. This table is searchable using criteria such as the year played, seasonality (regular season/playoffs), among others.

Following the assembly of the large multidimensional dataset, we designed statistical analysis modules to analyze player performance. Through correlation analysis, we constructed multidimensional variables to capture individual player performance, which in turn helped to rank players and teams. The procedure of correlation analysis will be described in more detailed later. Using statistical tests, we determined the most influential factors on game outcomes. These modules are illustrated in Figs. 3 and 5.

With the key factors identified, we designed algorithms to predict game outcomes. The design of the process is outlined in Fig. 6, and the methods of prediction are detailed in Section 4.3. Overall, data collection, key factor analysis, game analyses, player performance evaluation, and outcome prediction are the main functions of the system.

[2] Fenwick rating: unblocked shot attempts by NHL; named after blogger Matt Fenwick.

[3] Corsi rating: shot attempted differential (shots on goal, missed shots on goal and blocked shot attempts towards the opposition's net minus the same shot attempts directed at your own team's net) while at even-strength play.

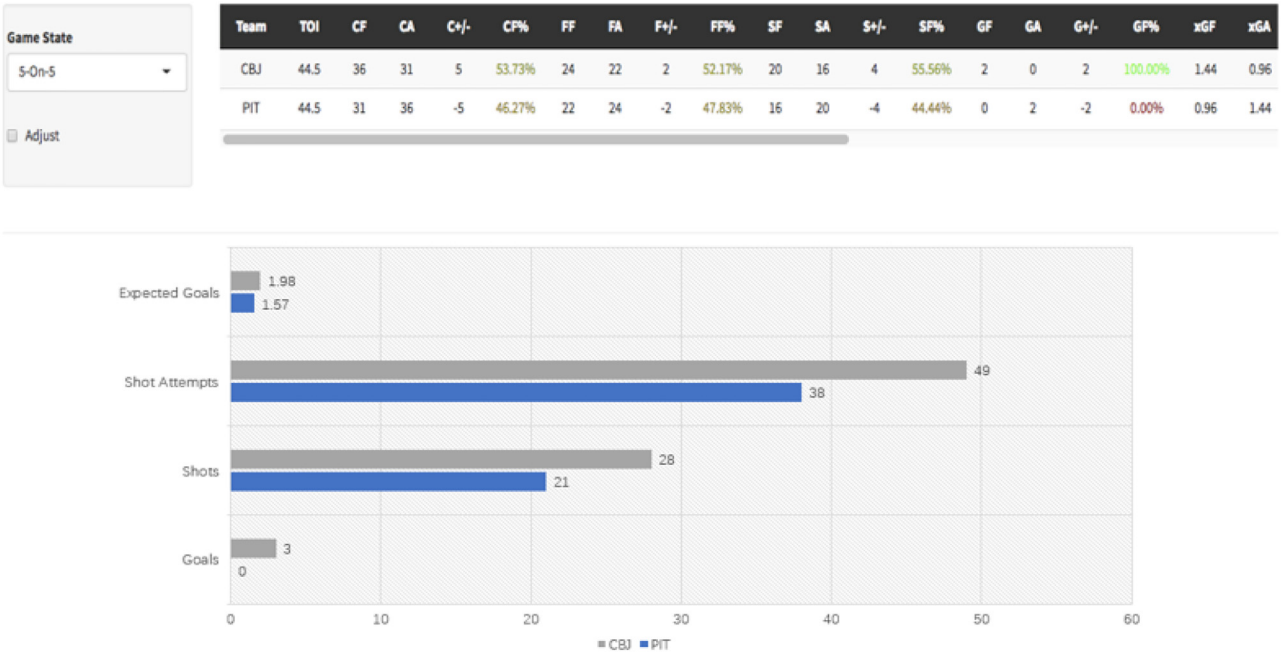

Fig. 2. Head-to-Head Team Performance Comparison.

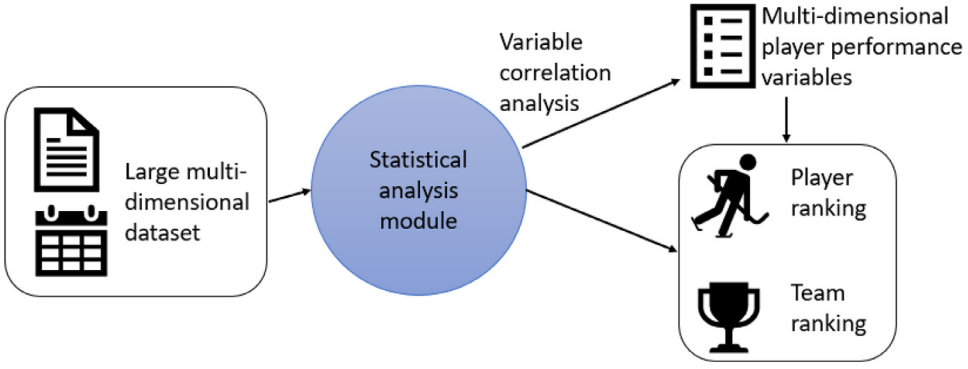

Fig. 3. Player Performance Analysis Module.

2016-2017 | Regular Season | 5v5 | All Scores | Individual | Counts | Submit
Filters [+]

Pittsburgh Penguins

Show/Hide Columns

	Player	Position	GP	TOI	Goals	Total Assists	First Assists	Second Assists	Total Points	Shots	SH%	iCF	iFF	iSCF	iHDCF	Rush Attempts	Rebounds Created	PIM	Total Penalties	Minor	Major	Misconduct
1	Sidney Crosby	C	75	1111:18	26	24	17	7	50	173	15.03	279	217	186	84	4	15	16	8	8	0	0
2	Patric Hornqvist	R	70	874:51	11	15	10	5	26	173	6.36	253	206	168	92	11	17	12	6	6	0	0
3	Phil Kessel	C	82	1097:51	14	23	13	10	37	158	8.86	263	197	130	44	8	14	8	4	4	0	0
4	Conor Sheary	L	61	830:54	16	26	10	16	42	126	12.7	206	165	143	67	7	18	16	8	8	0	0
5	Scott Wilson	C	78	808:46	8	18	11	7	26	121	6.61	205	163	119	61	4	15	22	8	6	2	0
6	Evgeni Malkin	C	62	865:10	20	21	17	4	41	115	17.39	181	150	134	62	3	12	52	25	25	0	0
7	Chris Kunitz	L	71	909:28	8	14	7	7	22	113	7.08	189	146	112	45	7	13	22	11	11	0	0
8	Justin Schultz	D	78	1296:16	7	16	7	9	23	112	6.25	215	147	81	11	3	12	24	12	12	0	0
9	Carl Hagelin	L	61	766:40	5	14	8	6	19	109	4.59	174	149	92	41	4	12	10	5	5	0	0
10	Nick Bonino	C	80	955:19	11	12	7	5	23	108	10.19	166	132	112	51	6	12	8	4	4	0	0
11	Bryan Rust	R	57	694:58	10	11	7	4	21	99	10.1	172	131	114	52	6	11	8	4	4	0	0
12	Kris Letang	D	41	737:54	2	12	2	10	14	87	2.3	154	114	52	7	2	6	16	8	8	0	0
13	Ian Cole	D	81	1352:13	4	18	8	10	22	86	4.65	212	151	56	9	0	14	40	18	18	0	0
14	Jake Guentzel	C	40	558:57	15	12	9	3	27	71	21.13	117	93	87	48	5	10	8	4	4	0	0
15	Matt Cullen	C	72	758:43	8	14	8	6	22	70	11.43	118	94	84	52	5	9	18	8	8	0	0
16	Brian Dumoulin	D	70	1202:48	0	8	4	4	8	68	0	162	91	53	10	4	10	10	5	5	0	0
17	Trevor Daley	D	56	861:31	4	8	2	6	12	67	5.97	122	91	51	15	1	4	18	9	9	0	0
18	Olli Maatta	D	55	833:36	1	6	4	2	7	63	1.59	122	84	35	4	1	3	10	5	5	0	0
19	Eric Fehr	C	52	446:42	5	5	3	2	10	51	9.8	82	67	55	24	2	5	8	4	4	0	0
20	Tom Kuhnhackl	R	57	494:45	2	10	5	5	12	42	4.76	86	62	49	20	3	6	18	9	9	0	0

Fig. 4. Sample Table for Player Performance Metrics.

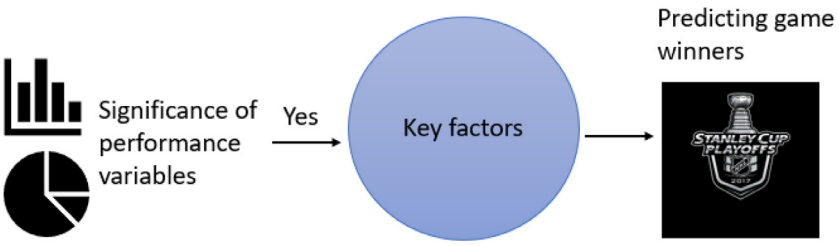

Fig. 5. Key Factor Analysis Module.

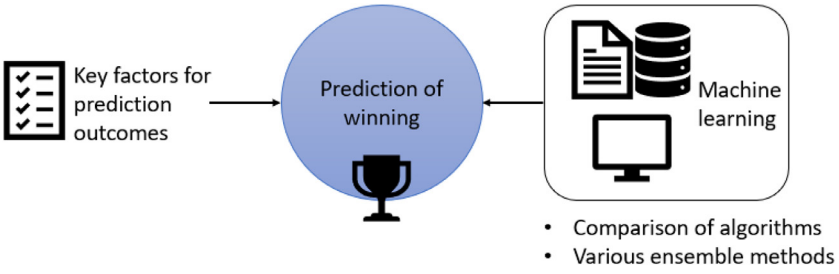

Fig. 6. Prediction of Game Outcomes.

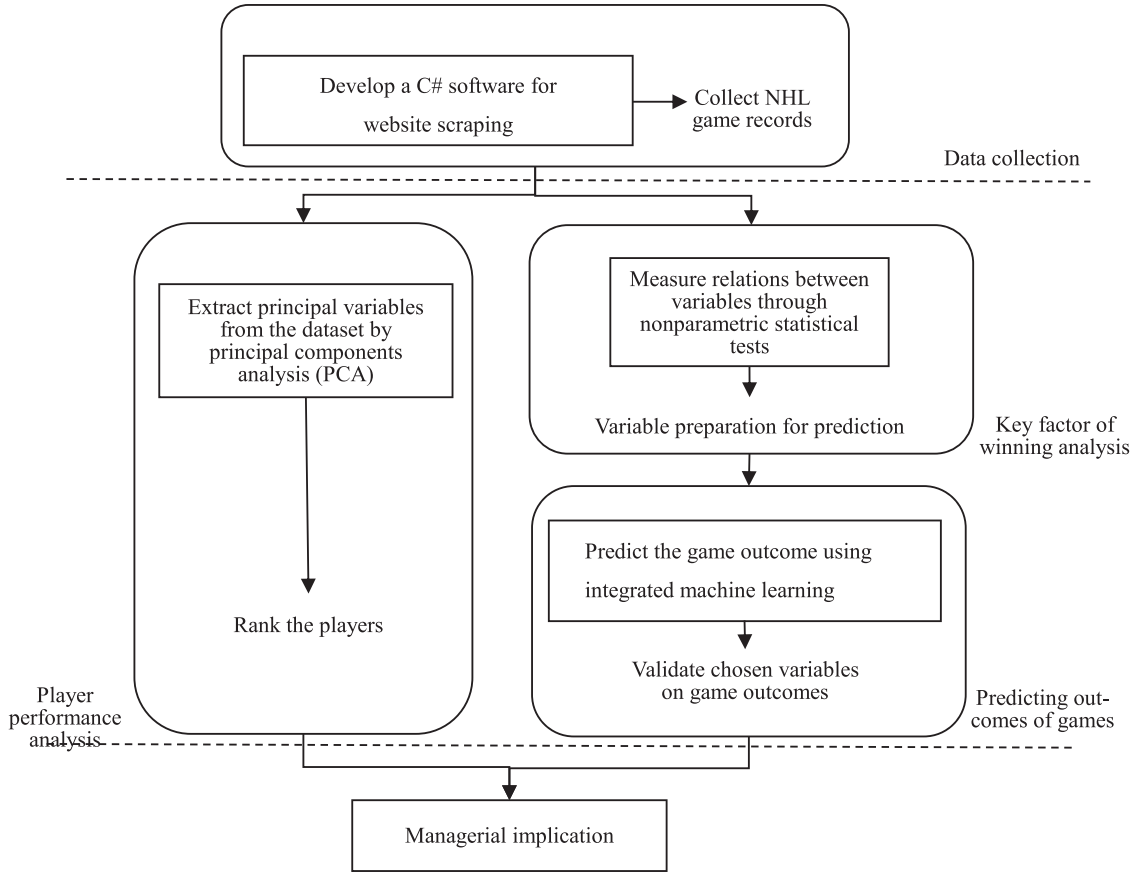

Fig. 7. Information Flow Among Key System Components.

Fig. 7 summarizes how we used a number of web scraping and machine learning techniques to generate player rankings and game prediction factors for managerial implications.

4. Methodology

We exemplify the methodologies used in the system with data from the 2014–15 season. During that season, 862 skaters and 92 goalies partook in a total of 1230 regular season games (2460 team records) and 89 post-season games (178 records). Table 2 lists the 18 variables we collected for each player in each game, with a total of more than 43 million observations: 2549 games $\times$ 18 variables $\times$ (862 skaters + 92 goalies). Alternatively, Table 10 defines the variables collected for each team per game, totaling 66,274 (2549×26) team observations.

Table 2
Variables Collected on Players for the 2014–15 Season.

No.	Item	Explanation
–	**PLAYER**	Player name
–	**TEAM**	Team name
–	**POS**	Position
1	**GP**	Total number of games played
2	**G**	Total number of goals*
3	**A**	Total number of assists
4	**+/–**	Plus/minus. When an even-strength or shorthanded goal is scored, every player on the ice from the goal-scoring team is credited with a "plus" and every player on the ice from the team scored against gets a "minus." A player's overall total is calculated by subtracting the minuses from the pluses.
5	**PIM**	Total number of penalty minutes
6	**PPP**	Total number of power play points
7	**SHP**	Total number of shorthanded points
8	**OT**	Total number of overtime goals
9	**S**	Total number of shots
10	**S%**	Shooting percentage
11	**TOI/GP**	Time on-ice per game
12	**SHIFT/GP**	Average shifts per game
13	**HITS**	Total number of hits
14	**BKS**	Total number of blocked shots
15	**MSS**	Total number of missed shots
16	**GVA**	Total number of giveaways
17	**TKA**	Total number of takeaways
18	**FO%**	Faceoff win percentage

* Goals scored during a shootout do not count toward a player's goal total.

4.1. Principal components analysis (PCA)

Player performance directly and significantly influences game outcomes. Yet in both academia and practice, player performance metrics have often been evaluated only one factor at a time. We address this shortcoming by combining and analyzing all factors collectively from a large, multidimensional dataset. As variables within the dataset could be highly correlated, we confirm suitability for employing PCA to reduce dimensions. We then extract the correlated variables into a representative set of linearly uncorrelated variables called the principal components.

Data validation consists of three phases. We first ran the Kaiser–Mayer–Olkin Test (*KMO*) to check for sampling adequacy, i.e., the appropriateness of using factor analysis. *KMO* takes a value between 0 and 1; larger values (between 0.5 and 1.0) indicate that factor analysis is appropriate (Tabachnick & Fidell, 2001). Second, we used Barlett's Test of Sphericity (p) to simplify the dataset by paring away redundant variables; we set a threshold of $p < 0.05$. Third, we checked the correlation matrix—if the off-diagonal values are high, then the variables are sufficiently correlated; if these values are close to zero, there is little correlation between the variables. When $KMO > 0.5$, $p < 0.05$, and the correlation index > 0.3, the data passes the tests.

Subsequent to data validation is application of PCA. The initial step of PCA generates the same number of components as variables (18, as defined in Table 2), sorted by the amount of explained variance. The proportion of variance explained by the i^{th} principal component is determined by dividing the eigenvalue of that component by the sum of all eigenvalues. A large proportion of the variance may be explained by a few principal components, reducing the dimensionality of the problem. For our analysis, we only retained components with eigenvalues ≥ 1.

We applied PCA to the collected data to rank the players and teams. After dividing players into skaters and goalies, we provided player rankings for the regular and postseasons. The remaining subsections summarize the process and results. Applicably, team management may use the resulting player standings to determine player compensation.

Table 3
Total Variance Explained for Regular Season Skaters.

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	8.414	46.747	46.747
2	2.145	11.917	58.664
3	1.589	8.825	67.490
4	1.051	5.838	73.328
5	0.897	4.986	78.313
6	0.793	4.407	82.720
7	0.700	3.891	86.612
8	0.605	3.362	89.974
9	0.419	2.329	92.303
10	0.356	1.979	94.282
11	0.271	1.508	95.790
12	0.221	1.227	97.017
13	0.195	1.082	98.099
14	0.130	0.724	98.822
15	0.098	0.546	99.369
16	0.049	0.270	99.639
17	0.037	0.204	99.843
18	0.028	0.157	100.000

Table 4
Component-Loading Matrix for Regular Season Skaters.

	Component			
	1	2	3	4
GP	0.870	−0.061	−0.309	−0.008
G	0.808	0.475	0.051	0.011
A	0.903	0.131	0.180	−0.075
+/–	0.209	0.090	0.349	0.709
PIM	−0.470	0.139	0.639	−0.061
PPP	0.790	0.233	0.281	−0.290
SHP	0.400	0.016	−0.130	0.611
OT	0.459	0.244	0.270	0.076
S	0.927	0.211	−0.002	−0.084
S%	0.319	0.469	−0.062	0.167
TOI/GP	0.762	−0.480	0.257	−0.053
Shift/GP	0.719	−0.506	0.234	−0.007
FO%	0.197	0.599	−0.358	−0.081
BkS	0.559	−0.695	−0.048	0.080
MsS	−0.920	−0.100	−0.006	0.104
GvA	−0.847	0.248	−0.098	0.092
TkA	0.839	0.211	0.020	0.023
Hits	0.501	−0.220	−0.703	0.069

*See Table 2 for the descriptions of each abbreviated variable.

4.1.1. Skater ranking based on regular season performance data

As *KMO* was 0.870 for the regular season skater data, and Bartlett's Test of Sphericity yielded $p < 0.01$, PCA was deemed suitable for the dataset.

PCA first generated 18 components, each of which is a linear combination of the original 18 variables. Table 3 details the total and cumulative variance explained by each component. The first four components account for 73.33% of the variation in players' performance scores, and thus were designated as the "key" components. The first four components, denoted by F_1, F_2, F_3 and F_4, were selected due to their high eigenvalues (≥ 1); the remaining components have successively less explanatory power.

Using the four key components, we proceeded from PCA to development of the component-loading matrix (see Table 4). The correlation of the variables with each component can be examined by the magnitude of the values in the matrix. For instance, *GP* (i.e., number of games played by a player) strongly correlates with the first component (0.870), but not with the other three components.

Table 5 presents the coefficients for computing the principal component scores.

Table 5
Component Score Coefficient Matrix for Regular Season Skaters.

	Component			
	F1	F2	F3	F4
GP	0.103	−0.029	−0.195	−0.008
G	0.096	0.222	0.032	0.011
A	0.107	0.061	0.114	−0.071
+/−	0.025	0.042	0.220	0.675
PIM	−0.056	0.065	0.402	−0.058
PPP	0.094	0.109	0.177	−0.276
SHP	0.047	0.008	−0.082	0.582
OT	0.055	0.114	0.170	0.072
S	0.110	0.098	−0.001	−0.080
S%	0.038	0.219	−0.039	0.159
TOI/GP	0.091	−0.224	0.162	−0.050
Shift/GP	0.085	−0.236	0.147	−0.007
FO%	0.023	0.279	−0.225	−0.077
BkS	0.066	−0.324	−0.030	0.076
MsS	−0.109	−0.047	−0.004	0.099
GvA	−0.101	0.116	−0.062	0.088
TkA	0.100	0.098	0.013	0.022
Hits	0.060	−0.103	−0.442	0.065

*See Table 2 for the descriptions of each abbreviated variable.

For example, the specification for first component F_1 is as follows:

$$\begin{aligned} F_1 = {} & 0.103 \times GP + 0.096 \times G + 0.107 \times A + 0.025 \times (+/-) \\ & - 0.056 \times PIM + 0.094 \times PPP + 0.047 \times SHP + 0.055 \times OT \\ & + 0.110 \times S + 0.038 \times S\% + 0.091 \times (TOI/GP) + 0.085 \\ & \times (Shift/GP) + 0.023 \times FO\% + 0.066 \times BkS - 0.109 \times MsS \\ & - 0.101 \times GvA + 0.100 \times TkA + 0.060 \times Hits \end{aligned} \quad (1)$$

We can then plug in player j's corresponding values to determine his component score, F_{1j}. The same approach is repeated to determine the other three component scores (F_{2j}, F_{3j}, F_{4j}) for the same player.

After simplifying the original 18 variables into four principal components that collectively explain the majority (73.3%) of variation in player performance, we weigh each principal component F_{ij} by the proportion of variance it explains, leading to each player's overall score. The total score for player j, Y_j, is calculated as:

$$Y_j = \frac{\lambda_1}{\sum \lambda} F_{1j} + \frac{\lambda_2}{\sum \lambda} F_{2j} + \frac{\lambda_3}{\sum \lambda} F_{3j} + \frac{\lambda_4}{\sum \lambda} F_{4j} \quad (2)$$

Using Eq. (2), we determined the composite score for each of the 882 skaters playing in the regular season. The top 20 skaters, according to our assessment approach, are displayed in Table 6 and Table 7.

4.1.2. Skater ranking based on post-season performance

In post-season, fewer teams play but each team faces the same opponent multiple times. The overtime rules differ from the regular season, as there are no shootouts during the playoffs. Thus, we consider post-season performance separately. Following the same procedure, we ranked the performances of the 339 skaters from the 16 teams participating in the post-season. Table 8 lists the top 10 skaters as ranked by post-season performance.

4.1.3. Ranking goalies based on regular season performance data

Likewise, we followed the same steps (KMO and Bartlett's Sphericity Test) to confirm suitability of PCA on the 92 goalies' performances. Table 8 presents the top 20 goalies resulting from this method.

Similarly, we ranked 24 goalies from the 16 playoff teams based on their performances in the post-season games. Table 9 lists the top 10 goalies identified with this method.

Table 6
Top 20 Skaters in the Regular Season via PCA Approach.

Rank	Player	Team	Score
1	Alex Ovechkin	WSH	162.55
2	John Tavares	NYI	160.06
3	Jamie Benn	DAL	157.74
4	Max Pacioretty	MTL	150.41
5	Rick Nash	NYR	145.67
6	Blake Wheeler	WPG	131.37
7	Ryan Getzlaf	ANA	120.97
8	Sean Monahan	CGY	117.36
9	Ryan Kesler	ANA	112.72
10	Jonathan Toews	CHI	111.45
11	Brad Marchand	BOS	110.34
12	Tyler Toffoli	LAK	109.92
13	Joe Pavelski	SJS	107.39
14	Filip Forsberg	NSH	104.67
15	Mark Stone	OTT	103.89
16	Erik Karlsson	OTT	103.11
17	Steven Stamkos	TBL	102.58
18	Logan Couture	SJS	102.42
19	Andrew Ladd	WPG	101.68
20	Nick Foligno	CBJ	101.62

Table 7
Top 10 Skaters in the Post-season via PCA Approach.

Rank	Player	Team	Score
1	Matt Beleskey	ANA	54.15
2	Brandon Bollig	CGY	22.51
3	Sam Bennett	CGY	14.89
4	Joakim Andersson	DET	11.47
5	Nicklas Backstrom	WSH	10.08
6	Mikael Backlund	CGY	8.01
7	Jay Beagle	WSH	6.70
8	Bryan Bickell	CHI	5.66
9	Patrik Berglund	STL	5.07
10	Beau Bennett	PIT	4.82

Table 8
Top 20 Goalies in the Regular Season via PCA Approach.

Rank	Player	Team	Score
1	Braden Holtby	STL	131.07
2	Jonathan Quick	NSH	129.88
3	Marc-Andre Fleury	STL	115.99
4	Carey Price	CGY	115.16
5	Cory Schneider	OTT	104.53
6	Kari Lehtonen	BOS	103.97
7	Tuukka Rask	NYI	99.69
8	Ben Bishop	CGY	98.23
9	Antti Niemi	DET	96.29
10	Pekka Rinne	NYR	94.70
11	Devan Dubnyk	CBJ	85.66
12	Jaroslav Halak	LAK	83.07
13	Semyon Varlamov	NYI	76.35
14	Steve Mason	NYR	74.12
15	Frederik Andersen	NYI	74.06
16	Corey Crawford	OTT	73.06
17	Mike Smith	TBL	71.82
18	Roberto Luongo	NYR	70.46
19	Jonathan Bernier	WPG	62.32
20	Sergei Bobrovsky	ANA	60.81

4.2. Nonparametric tests

We used nonparametric statistical methods to analyze the performance data and answer important questions about hockey games. Specifically, we employed Wilcoxon's rank-sum test, a nonparametric alternative to the two-sample t-test, to determine if two independent samples were from the same population; this delineates the significance of each variable on the outcome of winning or losing, and assists in selecting the proper metrics for game outcome prediction. As a result, significant variables are then designated as independent variables in the prediction model.

Table 9
Top 10 Goalies in the Post-season via PCA Approach.

Rank	Player	Team	Score
1	Ben Bishop	CGY	198.70
2	Henrik Lundqvist	NYR	85.41
3	Corey Crawford	OTT	82.39
4	Frederik Andersen	NSH	74.41
5	Braden Holtby	NYR	70.74
6	Carey Price	WSH	55.69
7	Devan Dubnyk	NYI	23.79
8	Petr Mrazek	ANA	12.30
9	Jaroslav Halak	CBJ	11.28
10	Craig Anderson	STL	1.544

Table 10
Variables Collected on Teams.

No.	Item	Explanation
–	**Team**	Team Name
–	**Season**	Season Type
–	**R/H**	Road Game or Home Game
–	**Date**	Game Date
1	**GF**	On-Ice Goals For Total
2	**GA**	On-Ice Goals Against Total
3	**Dec**	Decision of the Game
4	**SV%**	Save Percentage
5	**CF%**	Corsi%: The percentage of on-ice shot attempts (on goal, missed, or blocked)
6	**CF**	Corsi For Total
7	**CA**	Corsi Against Total
8	**FF%**	Fenwick For Percentage of Total
9	**FF**	Fenwick For total
10	**FA**	Fenwick Against Total
11	**MSF**	Missed Shots For
12	**MSA**	Missed Shots Against
13	**BSF**	Blocked Shots For
14	**BSA**	Blocked Shots Against
15	**SF%**	Shots on Goal For Percentage
16	**SF**	Shots on Goal For Total
17	**SA**	Shots on Goal Against Total
18	**Shoot%**	Shooting Percentage
19	**FO%**	Faceoff Winning Percentage
20	**FO_W**	Faceoffs Won
21	**FO_L**	Faceoffs Lost
22	**HIT**	Hits
23	**HIT-**	Hits Taken
24	**PN**	Penalties
25	**PN-**	Penalties Drawn
26	**PenD**	Penalty Differential

The rank-sum test is performed to compare each of the 26 metrics in Table 10 (explicating team performance history) for winning and losing teams. The test results are summarized in Table 11, in which "TRUE" indicates that the difference is statistically significant, consequently specifying the corresponding factors as significantly influential on the outcome of the game. Alternatively, "FALSE" indicates that the factor is not important. Finally, directions "+" and "–" symbolize that the factor has a positive or negative effect, respectively, on winning. Table 11 outlines the 19 factors that were found to be significant in predicting game results.

4.3. Machine learning for prediction

4.3.1. Support vector machine learning for predicting games

Support vector machine (SVM) is a supervised machine learning model, typically used for pattern recognition, classification, and regression analysis (Burges, 1998). We took the historical performance as input to predict the outcomes of hockey games. Using the 19 variables labeled as TRUE in Table 11, we applied the SVM technique to predict game outcomes. We also used SVM to confirm the appropriateness of the chosen set of variables for game prediction.

We build a SVM classifier from the historical game data set, $S=\{(X^i,\ y^i),\ i=1,\ 2,\ \cdots,\ m\}$, where $X^i=(x_1^i, x_2^i, \cdots, x_n^i)$, and $m$represents the number of instances ($m=1230$), and n is the number of input variables ($n=19$). y^i in our case is a binary categorical variable that 1 represents win and -1 represents lose. z represents our prediction for a game result which is a linear combination of the attributes (x_i) multiplied by corresponding weights (w_i), plus a noise term (b):

$$Z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b \tag{3}$$

X^i denotes the inputs of the 19 variables for game i. $Z=-1$ indicates a loss, while $Z=1$ indicates a win. Eq. (3) can be determined by the Sequential Minimal Optimization (SMO) algorithm (Platt, 1998), an algorithm used to solve SVM quickly. If $Z > 0$, the team being analyzed is expected to win the game; otherwise, the team is expected to lose. The training set is comprised of 1230 randomly selected game records, and the remaining 1230 records constitute the test set. After training the SMO, we were able to identify the weights for the SVM classifer (see Table 12).

We trained the SVM to learn from the 1230 regular season records by populating the metrics for each record, with team won or lost tags. Then we predict the game outcome on the training set and obtain 94.05% accuracy on the training set. The results suggested that the variables we selected are quite accurate in predicting NHL game outcomes, and the trained SVM classifier based on training set is valuable.

We then use the test set to determine whether the now-intelligent SVM could correctly predict a game based only on the historical performance.

4.3.2. Ensemble methods of machine learning for prediction

Classical machine learning searches within the hypothesis space, constructed based on all possibilities, for the classifier function $\boldsymbol{h}$ that best represents the actual conditions. Classical techniques include decision tress, artificial neural networks, naïve Bayes, and SVM, among others. A more integrated machine learning approach uses multiple learners under an ensemble rule—similar to a decision made by a committee—to generate a collective result superior to that of any single machine.

Dietterich (2000) has demonstrated the effectiveness of integration from the statistical, calculative, and representative standpoints:

(1) For general learning tasks, the initial search space is very large. However, if the number of training instances is too small, the assumptions and selected output-classifier of the algorithm will adequately suit the training set but fail in general practice. The integration of multiple hypotheses can reduce overfitting and improve generalizability of the models.
(2) Obtaining the best classifier is well-known as an NP-hard problem in artificial neural network learning and decision tree learning (Dietterrich, 1997). Other classifier models face similar computational complexity. Certain heuristic algorithms can simplify the search for the null hypothesis, although at the risk of generating suboptimal results. Integrating multiple hypotheses can generate a result closer to the actual objective function value.
(3) Since the assumption space is arbitrary, the actual objective assumptions are not in the assumed space for most machine learning applications. If we assume that the space is not closed under some ensemble calculation, then it is possible to represent the target assumptions that are not assumed in space by combining a set of assumptions in the hypothetical space.

Table 11
Impacts of Historical Performance on Game Outcome.

Factors	Direction	Whether affect	Factors	Direction	Whether affect
GF	+	TRUE	**SF%**	+	TRUE
GA	−	TRUE	**SF**	+	TRUE
G+/−	+	TRUE	**SA**	−	TRUE
SV%	+	TRUE	**Shoot%**	+	TRUE
CF%	−	TRUE	**FO%**	+	TRUE
CF	−	TRUE	*FO_W*	+	*FALSE*
CA	+	TRUE	*FO_L*	−	*FALSE*
FF%	+	*FALSE*	**HIT**	−	TRUE
FF	+	*FALSE*	**HIT-**	+	TRUE
FA	−	*FALSE*	*PN*	−	*FALSE*
MSF	−	TRUE	*PN-*	+	*FALSE*
MSA	+	TRUE	**Pen D**	+	TRUE
BSF	−	TRUE	**BSA**	+	TRUE

*Table 10 provides descriptions for each variable abbreviation.

Table 12
Weights Identified by SMO.

Metric	Weight	Metric	Weight
GF	0.7554	BSA	0.0078
GA	−0.6408	SF%	−0.0039
G+/−	1.3962	SF	0.0228
SV%	−0.0127	SA	0.0210
CF%	0.1109	Shoot%	−0.0204
CF	−0.0254	FO%	0.0301
CA	0.0200	HIT	0.0167
MSF	0.0077	HIT-	−0.0234
MSA	−0.0088	PenD	−0.2026
BSF	−0.0559		

*Table 10 provides descriptions for each variable abbreviation.

To compare the performance of the learners, we applied various ensemble methods: Bagging, AdaBoost, Boosting, and RobustBoost. In the Bagging algorithm, the training set of each classifier is composed of several examples randomly selected from the original training set. The training set usually has the same size as the original training set; because training case selection is replaceable, some instances of the original training set may appear multiple times in the new training set while other instances may not appear at all. The Bagging method improves the variability of the neural network by reselecting the training set, thus enhancing its generalization capability. The sensitivity of the result on the training set relates to the efficacy of Bagging. Bagging can improve the prediction accuracy of unstable, high-sensitivity learning algorithms (e.g., decision trees and neural networks), but not stable, low-sensitivity learning algorithms (e.g., *k*-nearest neighbor), and may even reduce the prediction accuracy of stable learning algorithms.

Unlike Bagging, Boosting features nonrandom, dependent training set selection; each iteration of Boosting training is related to the previous round of learning results. Furthermore, with Boosting, predictive functions are weighted and generated sequentially. For learning algorithms like neural networks, Bagging saves time due to its parallel and unweighted training.

In our algorithm, 2000 and 638 games were allocated as training and testing samples, respectively. The results are compared in Table 13.

Table 13 indicates that ensemble classifiers outperform individual classifiers. In our research, the prediction is based on Bagging, AdaBoost, Boosting, and RobustBoost strategies. All of the ensemble classifiers except KNN outperformed the previously reported SVM model, with greater than 90% prediction accuracy. We found that discriminant analysis with RobustBoost performed best on the testing set (91.8% accuracy) among the ensemble classifiers we considered. When employing either discriminant analysis or decision trees, we observed a preference for Boosting ensemble strategies over Bagging. However, discriminant analysis always performed better than decision trees, given the same ensemble strategy. While the preferable ensemble strategy varies by classifier, our results indicate that implementing one of these strategies is beneficial for test set prediction.

5. Managerial implications

Since the NHL imposes a salary cap on each team, team management is faced with capital allocation questions. Are star skaters worthy of a multimillion dollar salary, which often is disproportionate to the salaries of his teammates? Should an effective goalie be compensated significantly higher for his contribution to wins?

Hughes and Bartlett (2002) studied various performance metrics in different sports and offer recommendations on how to use these performance metrics for player evalution and team management. We used the skater and goalie ratings developed in Tables 5 and 6 to examine the roles of star skaters and goalies in winning games.

5.1. Star skaters' roles

In Table 6, we used Eq. (2) to rank all players. For this analysis, we used the Spearman rank correlation to examine the relationship between the league ranking of each team's best player and his team's standing at the end of regular season. We found the correlation to be 0.353 and significant, indicating that star skaters indeed influence game outcomes. However, because the correlation coefficient is a low 0.353, a team's star player does not dominate the team's overall performance.

We infer from this result that recruiting a single star player may not be the best way to increase wins. An NHL team may be more benefited by investing in excellent teamwork and balance rather than over-emphasizing a single star player. We acknowledge that players' salaries are negotiated through multi-year contracts, so average salaries over the contract period cannot be determined by a single season's superior (or inferior) performance. However, the intra-team spread of player salaries can be considered in this context. Some teams compensate the top two to five skaters similarly. However, current publicly available data suggests that the highest paid skater on many NHL teams earns millions more annually than the next highest paid skater on his team. As might be expected with multi-year contracts, the highest paid skaters for each team do not necessarily align with the highest single-season rankings from our PCA approach. While our approach cannot include every possible factor contributing to salary, the PCA approach may benefit team management by providing additional insights, especially when drafting players and negotiating new contracts.

Table 13
Comparison of Ensemble Algorithm Results.

Classifier	Ensemble Strategy	Accuracy of Training Samples (%)	Accuracy of Testing Samples (%)
KNN	None	100.0000	80.4075
SVM	None	95.1500	90.5956
Bayes	None	94.5500	90.4389
Discriminant	None	95.0000	91.5361
Discriminant	Adaboost	94.6000	91.5361
Tree	Adaboost	96.6500	90.4389
Discriminant	Bagging	95.0000	91.0658
Tree	Bagging	100.0000	90.1254
Tree	Boosting	98.8000	91.3793
Discriminant	RobustBoost	95.5000	91.8495
Tree	RobustBoost	96.8500	90.4389

* Accuracy = (N/M) × 100 where N is the number of accurate classifications and M is the number of testing samples.

5.2. Goalies' roles

Espina-Agulló, Pérez-Turpin, Jiménez-Olmedo, Penichet-Tomás, and Pueo (2016) discuss the importance of male handball goalkeepers; however, no research has been conducted to examine the importance of goalies' role in ice hockey. In this paper, through Spearman rank correlation, we identify the relationship between the league ranking of a team's best goalie and his team's standing. We find the correlation is 0.583, which is higher than that of the star skaters and thus warrants further examination.

Save percentage (*SV%* in Table 10) is the proportion of shots on goal that a goalie was able to stop. To check if a capable goalie (one with a high *SV%*) helps win the game, we conducted a rank-sum test to compare the save percentages from all games won to those from all games lost. We rejected the null hypothesis ($Z = -29.18$, $p < 0.001$), which indicates that the *SV%* of the winning and losing teams differ significantly, and consequently *SV%* significantly impacts the game outcome. The mean *SV%* for the winning and losing teams was 93.98% and 87.74%, respectively. Thus, goalies play a critical role in the game, and team management can justify higher compensation for an excellent goalie.

This result contradicts with the current practice of most teams. Based on publicly available NHL salary data (average salary per year over length of contract) for the 2015–16 season, the star goalie is not typically the highest paid member on the team. In fact, only four NHL goalies are the highest paid players on their respective teams. Like the star skaters, when the salaries of all goalies across the NHL are compared, the ordering is not consistent with our PCA-based rankings. As previously stated, salary contracts span multiple years, and thus average salary rankings are not expected to closely reflect single-season performance rankings. Other factors, such as differences in salary caps for each team, also play a part in this discrepancy. Still, the PCA approach to ranking goalies may provide management with a different perspective when negotiating contracts.

5.3. Differences between home and road games

Inspired by Vaz, Carreras, and Kraak (2012), who examined the impact of alternating home and away field advantage on rugby game performance, we examined the differences between home and road games as well as regular and post-season games.

Do teams perform better during home games or road games? Jones (2009) analyzed home scoring advantage in NHL games. Using a rank-sum test, we compared the win percentages for all home games to those of all road games. We rejected the null hypothesis ($Z = -4.56$, $p < 0.001$), which indicates that the average win percentages of home and away games differ significantly, and consequently game location affects the game outcome. Therefore, identifying which variables contribute to this difference is an important investigation. We again conducted rank-sum tests for each of the 26 team performance metrics and summarize the results in Table 14. Evidently, the performances vary by location, except for *SV%* ($p = 0.16$) and Shoot% ($p = 0.16$). Thus, location indeed significantly influences game performance, and there is a home game advantage—i.e., teams typically perform better on their own ice. Accordingly, the coaching staff can use the same approach to plan how team and player rotation should be adjusted on home-ice versus away.

Table 14
Rank-Sum Test Comparing Team Performance for Home and Road Games.

Factors	H	Z val	Rank sum	P
GF	TRUE	−3.5971	1.67E + 06	0.000322
GA	TRUE	3.5971	1.81E + 06	0.000322
G +/−	TRUE	−5.2209	1.64E + 06	<0.000001
SV%	FALSE	−1.417	1.71E + 06	0.156484
CF%	TRUE	−8.9891	1.56E + 06	<0.000001
CF	TRUE	−7.3237	1.60E + 06	<0.000001
CA	TRUE	7.3237	1.88E + 06	<0.000001
FF%	TRUE	−9.9533	1.55E + 06	<0.000001
FF	TRUE	−8.1608	1.58E + 06	<0.000001
FA	TRUE	8.1608	1.90E + 06	<0.000001
MSF	TRUE	−6.1692	1.62E + 60	<0.000001
MSA	TRUE	6.1692	1.86E + 06	<0.000001
BSF	TRUE	−2.8326	1.69E + 06	0.004617
BSA	TRUE	2.8326	1.80E + 06	0.004617
SF%	TRUE	−7.7046	1.59E + 06	<0.000001
SF	TRUE	−6.0239	1.62E + 06	<0.000001
SA	TRUE	6.0239	1.86E + 06	<0.000001
Shoot%	FALSE	−1.417	1.71E + 06	0.156484
FO%	TRUE	−13.9729	1.47E + 06	<0.000001
FO_W	TRUE	−10.7119	1.53E + 06	<0.000001
FO_L	TRUE	10.7164	1.95E + 06	<0.000001
HIT	TRUE	−4.405	1.65E + 06	0.000011
HIT-	TRUE	4.405	1.83E + 06	0.000011
PN	TRUE	3.0554	1.80E + 06	0.002247
PN-	TRUE	−3.0554	1.68E +06	0.002247
Pen D	TRUE	−6.126	1.62E +06	<0.000001

* Table 10 provides descriptions for variable abbreviations.

5.4. Differences between regular season and post-season games

Due to differences in the overtime rules, schedules, and number of teams playing, we treat the regular and post-season games differently. We used the Mann-Whitney U test as a method of rank-sum test to examine team performance differences between the regular and post-season games. Table 15 displays the results, which demonstrate that eight performance metrics differ significantly by season segment, i.e., CF, CA, BSF, BSA, FO_W, FO_L HIT,

Table 15
Rank-Sum Test Comparing Regular Season and Post-season Team Performance.

Factors	H	Z val	Rank sum	P
GF	FALSE	−1.414	205,307	0.157
GA	FALSE	−1.414	205,307	0.157
G+/−	FALSE	0.000	218,940	1.000
SV%	FALSE	1.488	233,536.5	0.137
CF%	FALSE	0.000	218,940	1.000
CF	***TRUE***	**2.082**	**239,363**	**0.037**
CA	***TRUE***	**2.082**	**239,363**	**0.037**
FF%	FALSE	0.000	218,940	1.000
FF	FALSE	0.084	219,759	0.933
FA	FALSE	0.084	219,759	0.933
MSF	FALSE	0.459	223,429	0.646
MSA	FALSE	0.459	223,429	0.646
BSF	***TRUE***	**4.027**	**258,383**	**0.000**
BSA	***TRUE***	**4.027**	**258,383**	**0.000**
SF%	FALSE	0.000	218,940	1.000
SF	FALSE	−0.273	216,260	0.785
SA	FALSE	−0.273	216,260	0.785
Shoot%	FALSE	−1.488	204,343	0.137
FO%	FALSE	−0.105	217,905	0.916
FO_W	***TRUE***	**2.232**	**240,817.5**	**0.026**
FO_L	***TRUE***	**2.402**	**242,481**	**0.016**
HIT	***TRUE***	**9.899**	**316,031**	**0.000**
HIT-	***TRUE***	**9.899**	**316,031**	**0.000**
PN	FALSE	−1.733	202,168	0.083
PN-	FALSE	−1.733	202,168	0.083
PenD	FALSE	0.000	218,940	1.000

* Table 10 provides descriptions for each variable abbreviation.

and HIT-. Table 16 compares the means of these eight metrics between the regular and post-season. All the metrics in the post-season have higher means, indicating that post-season games are of higher quality and more competitive. Therefore, their CF, CA, BSF, BSA, FO_W, FO_L HIT, and HIT- metrics are higher, but fewer penalty infractions occur. This result is consistent with common beliefs about the increased level of competition in playoffs.

5.5. Implication of machine learning prediction

Based on the analysis of team performance from the 2015 data, we found that our proposed approach can achieve high accuracy in predicting NHL game outcome (94.05% precision). This is a vast improvement relative to those in the literature. For example, Weissbock, Viktor, and Inkpen (2013) reached 59.8% accuracy when using SVM to predict hockey, which was later improved to 60.25% (Weissbock & Inkpen, 2014). For elite female hockey game prediction, Morgan et al. (2013) attained 64.3% accuracy when using machine learning. In comparison, Huang and Chang (2010) obtained 76.9% accuracy when predicting soccer games using neural networks, while Kahn (2003) achieved 75% accuracy for football game prediction. Similarly, Loeffelholz et al. (2009) attained 74.33% accuracy for basketball games. Finally, Miljkovic et al. (2010) used Naïve Bayes to predict NBA basketball games with 67% accuracy.

In our approach, we have taken full advantage of the historical information, carefully evaluated all possible indicators, and validated their efficacy. Compared with the most popular approach for game prediction, i.e., applying machine learning to individual judgment and textual contents (Pischedda, 2014; Weissbock & Inkpen, 2014; Weissbock et al., 2013), our approach is more objective and less dependent on subjective opinions. Using these results, team management can improve the number of wins by highlighting productive players, and similarly identify those who need improvement. Shapiro, DeSchriver, and Rascher (2012) show that a team's recent win-percentage history could significantly affect the prices of luxury suites in major North American sports facilities. This is one aspect of an organization's revenue, and thus, we believe that an increase in the number of wins will result in increased revenue for the team.

6. Conclusions

6.1. Summary and contribution

In this paper, we proposed an expert system with the integrated functionalities of collecting data; analyzing player and team performance; identifying factors important to winning; and predicting game outcomes. We developed a software module to collect NHL game records from a number of data sources. Subsequently, using performance data from the 2014–15 NHL season, we conducted PCA to develop composite rankings for players (skaters and goalies) and teams (regular and post-season games).

The contributions of our research are three-fold:

1. Development of an expert system for sport game outcome prediction
   Hockey games are highly stochastic, but their underlying operations and dynamics may be approximated by certain rules and logic. In the expert system, we collect historical data of hockey games, treat collected data, employ predictive models, and assess the models to ensure high quality outcome prediction. The expert system combines the intelligence and capability of an expert and can predict results accurately.
2. Incorporation of machine learning and big data in the expert system
   Our approach differs from existing single-metric ranking methods, as we use big data and machine learning techniques to both rank players by integrating various player performance metrics and evaluate teams by considering 26 team performance metrics. Nonparametric hypothesis tests are applied to select the metrics affecting game outcomes. The "TRUE" metrics detected are used as inputs for machine learning techniques.
3. Use of ensemble approach to derive highly accurate results

The high prediction accuracy (>90%) confirms that the SVM and ensemble machine learning algorithm are valuable tools that can accurately predict game outcomes. By employing machine learning on the training set, we also validated the use of the 19 "TRUE" metrics. The proposed expert system can help management improve team performance and make player recruitment and salary decisions. Furthermore, we utilized nonparametric methods to answer managerial questions. Given low explanatory power, we conclude that the star players—the primary focus of fans—are not by themselves the most important determinant of a game's outcome. Hockey requires teamwork that involves frequent player substitution, and our use of factors such TOI can more accurately capture on-ice performance. We also found that goalies play a significant role, as their save percentages exhibit a strong relationship with winning. The location of the game impacts the outcome and supports the notion of home-ice advantage. Similarly, there is a difference between the regular and post-season, as teams participating in the playoffs systematically demonstrate superior skills and face a greater level of competition.

6.2. Limitations & future work

While our proposed model predicts the outcomes of NHL games in a single season, the model is limited by the ever-changing nature of professional sports. The model should not be applied beyond a single season due to changes in players, coaching staff, and team management that may occur during the off-season. While changes to the roster or coaching staff may happen during the season, most long-term or large-scale changes occur in-between seasons. For example, a star player may change teams after deciding

Table 16
Comparing Regular Season and Post-season Game Performance.

	CF	CA	BSF	BSA	FO_W	FO_L	HIT	HIT-
Post-season	58.713	58.713	16.539	16.539	32.022	32.090	33.365	33.365
Regular-season	55.992	55.992	14.469	14.469	30.638	30.639	24.989	24.989

* Table 10 provides descriptions for each variable abbreviation.

not to renew the contract with his current team, or management may opt to replace the team's coach. Therefore, the model trained with 2014–2015 data would not be expected to perform well if applied to data from the 2015–2016 season. Similarly, we treated the regular season and post-season games separately because overtime rules differ during the playoffs. Furthermore, in-season disruptions to the roster, such as injuries, sometimes occur, and users of the system should be cautious of predictions under these circumstances.

While our proposed method performed well, additional data could further improve predictions. To achieve high accuracy for pre-game and in-game predictions, identification of historical performance metrics relevant to the game outcome is critical to incorporate in our system input. In the future, we could further explore the performance of the classifiers when different (or additional) input variables are used. For instance, the inclusion of player rankings and player metrics (individual or aggregate) might provide better insights into the link between player and team performance. Future research could also shed light on how moving-averaged and/or exponentially-smoothed team and player performance metrics can serve as our model inputs in a more dynamic setting. The involvement of expert judgments on intangible factors, such as coaches' tactics or a player's physical or mental assessment, is also desirable. The inclusion of coach data may improve the results since strategic and tactic decisions invariably affect game outcomes. Additionally, psychological factors are crucial in the game, since favorable mentality will improve attention, confidence, and strain capacity. Building such a multi-criteria, decision-making model, through the incorporation of robust input metrics and both tangible and intangible factors, could not only further improve game-predicting accuracy but also allow for a more fine-grained prediction interval.

Our analysis focuses on predicting wins and losses during a single NHL season, and future work could expand upon this. For instance, similar methods could be employed to predict whether or not a team will qualify for the playoffs or which team will win the championship. Formulating such predictions from a model trained with regular season data could become a factor in sports betting.

Conflict of interest

There is no conflict of interest.

Credit authorship contribution statement

Wei Gu: Conceptualization, Data curation, Formal analysis. **Krista Foster:** Writing - original draft, Writing - review & editing. **Jennifer Shang:** Supervision, Writing - original draft, Writing - review & editing. **Lirong Wei:** Visualization, Writing - review & editing.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (grant number 71702009, 71531013, 71729001), Beijing Social Science Foundation (reference no. 17JDGLA010, 18JDGLB034).

References

Andersson, P., Edman, J., & Ekman, M. (2005). Predicting the World Cup 2002 in soccer: Performance and confidence of experts and non-experts. *International Journal of Forecasting, 21*(3), 565–576.
Andersson, P., Memmert, D., & Popowicz, E. (2009). Forecasting outcomes of the World Cup 2006 in football: Performance and confidence of bettors and laypeople. *Psychology of Sport and Exercise, 10*(1), 116–123.
Aslan, B. G., & Inceoglu, M. M. (2007, October). A comparative study on neural network based soccer result prediction. In *In* Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on (pp. 545–550). IEEE.
Barry, D., & Hartigan, J. A. (1993). Choice models for predicting divisional winners in Major League Baseball. *Journal of the American Statistical Association, 88*(423), 766.
Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.
Buttrey, S. E., Washburn, A. R., & Price, W. L. (2011). Estimating NHL scoring rates. *Journal of Quantitative Analysis in Sports, 7*(3).
Carlin, B. P. (1996). Improved NCAA basketball tournament modeling via point spread and team strength information. *The American Statistician, 50*(1), 39–43.
Coleman, B. J. (2017). Team Travel Effects and the College Football Betting Market. *Journal of Sports Economics, 18*(4), 388–425.
Dietterrich, T. G. (1997). Machine learning research: Four current directions. *Artificial Intelligence Magazine, 4*, 97–136.
Dietterich, T. G. (2000). Ensemble methods in machine learning. In: *International workshop on multiple classifier systems* (pp. 1–15). Berlin, Heidelberg: Springer.
Dijksterhuis, A., Bos, M. W., van der Leij, A., & van Baaren, R. B. (2009). Predicting soccer matches after unconscious and conscious thought as a function of expertise. *Psychol Sci, 20*(11), 1381–1387.
Espina-Agulló, J. J., Pérez-Turpin, J. A., Jiménez-Olmedo, J. M., Penichet-Tomás, A., & Pueo, B. (2016). Effectiveness of male handball goalkeepers: A historical overview 1982-2012. *International Journal of Performance Analysis in Sport, 16*(1), 143–156.
Feltz, D. L., & Lirgg, C. D. (1998). Perceived team and player efficacy in hockey. *Journal of Applied Psychology, 83*(4), 557–564.
Huang, K. Y., & Chang, W. L. (2010). A neural network method for prediction of 2006 World Cup Football game. *The 2010 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.
Huang, K., & Chen, K. (2011). Multilayer perceptron for prediction of 2006 world cup football game. *Advances in Artificial Neural Systems, 2011*, 1–8.
Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of sports sciences, 20*(10), 739–754.
Jones, M. B. (2009). Scoring first and home advantage in the NHL. *International Journal of Performance Analysis in Sport, 9*(3), 320–331.
Kahn, J. (2003). Neural network prediction of NFL football games. *World Wide Web electronic publication* (pp. 9–15).
Kain, K. J., & Logan, T. D. (2014). Are sports betting markets prediction markets? Evidence from a new test. *Journal of Sports Economics, 15*(1), 45–63.
Leard, B., & Doyle, J. M. (2011). The effect of home advantage, momentum, and fighting on winning in the National Hockey League. *Journal of Sports Economics, 12*(5), 538–560.
Loeffelholz, B., Bednar, E., & Bauer, K. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports, 5*(1), 7.
Macdonald, B. (2012). An expected goals model for evaluating NHL teams and players. Paper presented at the. *MIT Sloan Sports Analytics Conference 2012*.
Miljkovic, D., Gajic, L., Kovacevic, A., & Konjovic, Z. (2010). The use of data mining for basketball matches outcomes prediction. Paper presented at the. *IEEE 2010 8th International Symposium on Intelligent Systems and Informatics (SISY)*.
Morgan, S., Williams, M. D., & Barnes, C. (2013). Applying decision tree induction for identification of important attributes in one-versus-one player interactions: A hockey exemplar. *J Sports Sci, 31*(10), 1031–1037.
Perlini, A. H., & Halverson, T. R. (2006). Emotional intelligence in the National Hockey League. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 38*(2), 109–119.
Platt, J. C. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines*. Microsoft Research.
Pischedda, G. (2014). Predicting NHL match outcomes with ML models. *International Journal of Computer Applications, 101*(9), 15–22.
Rimler, M. S., Song, S., & David, T. Y. (2010). Estimating production efficiency in men's NCAA college basketball: A bayesian approach. *Journal of Sports Economics, 11*(3), 287–315.
Shapiro, S. L., DeSchriver, T., & Rascher, D. A. (2012). Factors affecting the price of luxury suites in Major North American Sports Facilities. *Journal of Sport Management, 26*(3), 249–257.

Tarter, B., Kirisci, L., Tarter, R., Jamnik, V., Gledhill, N., & McGuire, E. J. (2009). Use of the Sports Performance Index for Hockey (SPI-H) to predict NHL player value. *International Journal of Performance Analysis in Sport, 9*(2), 238–244.

Tabachnick, B. G., & Fidell, L. S. (2001). Principal components and factor analysis. In *In using multivariate statistics* (pp. 582–633). Needham Heights, MA: Allyn & Bacon.

Van Roon, M. (2012). *Predicting the outcome of NBA playoffs*. Amsterdam: Vrije Universiteit.

Vaz, L., Carreras, D., & Kraak, W. (2012). Analysis of the effect of alternating home and away field advantage during the Six Nations Rugby Championship. *International Journal of Performance Analysis in Sport, 12*(3), 593–607.

Von Allmen, P., Leeds, M., & Malakorn, J. (2015). Victims or Beneficiaries?: Wage Premia and National Origin in the National Hockey League. *Journal of Sport Management, 29*(6), 633–641.

Voyer, D., & Wright, E. F. (1998). Predictors of performance in the National Hockey League. *Journal of Sport Behavior, 21*(4), 456–473.

Weissbock, J., & Inkpen, D. (2014). Combining textual pre-game reports and statistical data for predicting success in the national hockey league. In *In* Canadian Conference on Artificial Intelligence (pp. 251–262). Springer International Publishing.

Weissbock, J., Viktor, H., & Inkpen, D. (2013). Use of Performance Metrics to Forecast Success in the National Hockey League. Paper presented at the. *Workshop on Sports Data Mining at ECML/PKDD 2013*.

Yang, J. B., & Lu, C. H. (2012). Predicting NBA championship by learning from history data. In *Proceedings of artificial intelligence and machine learning for engineering design*.

Zimmermann, A., Moorthy, S., & Shi, Z. (2013). Predicting college basketball match outcomes using machine learning techniques: Some results and lessons learned. arXiv:1310.3607.