Journal of Sports and Games
Volume 3, Issue 2, 2021, PP 1-13
ISSN 2642-8466
DOI: https://doi.org/10.22259/2642-8466.0302001

Effects of the Use of Sports Analytics and Team Attributes on Success in Regular Season of National Hockey League

David Chu*and Gurdeepak Sidhu

Department of Mathematics and Statistics, University of the Fraser Valley, 33844 King Road, Abbotsford, BC Canada V2S 7M8

***Corresponding Author:** David Chu, Associate Professor, Department of Mathematics and Statistics, University of the Fraser Valley, 33844 King Road, Abbotsford, BC Canada V2S 7M8*

ABSTRACT

In this paper, we study the effects of the use of sports analytics and team attributes on teams' success in the regular season of the National Hockey League. A team's belief in analytics, the number of analytics staff, and the number of professional staff hired are examined for the use of sports analytics. Some of the team attributes considered here are the average age of players in a team, payrolls of different positions (goalies, defensemen, forwards), and numbers of the first-round draft picks in the previous three years. We shall examine the empirical data of 2014-2019 seasons. The team payroll is shown to be significantly positively correlated with a team's success in the regular season. It is interesting to see that teams scored 96 points or more are very likely advancing to playoffs, whereas teams scored 92 points or less are very unlikely advancing to playoffs. Four commonly used predictive modeling techniques (decision trees, random forests, logistic regressions, and neural networks) are applied to the data for classifying teams into playoffs or no playoffs. Random forests appear to be the best or as good as the other three techniques to yield the lowest validation misclassification error rate.

Keywords: *Team payroll, decision trees, random forests, logistic regressions, neural networks*

INTRODUCTION

In recent years, sports analytics has been very popular in the National Hockey League (NHL) among other major professional sports in North America. Chu and Wang (2019) examined empirical data to study the relationship between sports analytics and success in the regular season and postseason in Major League Baseball (MLB). As the NHL is very competitive these days, some managements would like to employ sports analytics to improve their team performance and have advantages over their competitors. They hope that it would increase their chances of winning more games for advancing to playoffs and might eventually generate better chances of winning a championship. Other managements, however, might be skeptical about the idea of sports analytics or might not even buy into the idea at all. Thus a few or even no analytics staff are hired in these teams.

In this paper, we study the effects of the use of sports analytics and various team attributes on teams' success in the NHL. To measure the use of sports analytics in NHL teams, we examine three aspects. The first aspect comes from a source outside of the teams. The analytics belief of NHL teams could be classified into five categories according to The Great Analytics Rankings (2015). These five categories were All-In, Believers, One-Foot-In, Skeptics, and Non-Believers. The second and third aspects come from some sources inside of the teams. We look at the number of analytics staff hired and dedicated to the analytics department, if any, for each team. Furthermore, we search for the total number of professional staff employed in each team. The professional staff include analytics staff, sports psychologists, nutritionists, speed skating coaches, etc. The categories of analytics belief and these numbers of analytics staff and professional staff might reflect teams' commitment to analytics and their support for improvement of teams' competence. We will assess how the use of sports analytics affects teams' success in the regular season in terms of their chances of advancing to the postseason. We hope in another paper to study how the use of sports analytics would affect teams' success in the postseason in terms of their stages towards the championship of Stanley Cup.

In addition to the factors of categories of analytics belief, the number of analytics staff, and the number of professional staff, we consider various team attributes that might have a positive impact on the outcomes of teams in the regular season. Some of these team attributes are the average age of players in a team, payrolls of different positions (goalies, defensemen, forwards), and numbers of the first-round draft picks in the previous three years. The complete list of these attributes is displayed in Section 2.

In Section 3, the conditional probabilities of teams' success in the regular season are estimated based on analytics belief, the number of analytics staff, and the number of professional staff. The Pearson sample correlation coefficients between various pairs of attributes are computed as well. The binary logistic regression model is explored to study the relationship of analytics belief, the number of analytics staff, and the number of professional staff with teams' success in the regular season. The team payroll in the NHL is shown to be significantly positively correlated with a team's success in the regular season. Schwartz and Zarrow (2009) showed a similar result that the team payroll in the MLB had a great positive influence on a team's success in the regular season.

Decision trees, random forests, logistic regressions, and neural networks are four commonly used predictive modeling techniques in data mining. They will be used in Section 4 to classify NHL teams into playoffs or no playoffs in the regular season. Three types of classification situations are considered for decision trees. Decision trees are sought based on (1) all team attributes available, (2) all team attributes excluding the number of regular-season wins, and (3) all team attributes excluding the number of regular-season wins and total points acquired as these two attributes are not available at the beginning of a season. Comparisons of these four techniques will be made as well with respect to the validation misclassification error rate. Finally, the conclusion and comments are given in Section 5.

NHL DATA SET

The data set we compiled here consists of records and attributes of NHL teams from the seasons of 2014-2019, i.e., beginning with the 2013/14 season and ending with the 2018/19 season. The attributes include the information on teams' categories of analytics belief as listed in The Great Analytics Rankings. The data set also includes the number of analytics staff and the number of professional staff hired by teams that can be tracked down from the Official NHL Record Book. Additionally, the regular-season wins, overtime losses, total points, the average age of players, and payrolls by positions (goalies, defensemen, forwards) of NHL teams for 2014-2019 are incorporated. These can be collected from the corresponding websites listed in the references. Also, the data set contains information on the competency of teams with regards to the goalies grade, defensemen grade, and forwards grade. All these attributes were assessed and given by The Hockey News Magazine.

Information on teams' first-round draft picks is also compiled to see if this would influence the future success of rebuilding teams. These attributes are the number of first-round picks, the highest pick, the number of first-round picks signed, numbers of first-round picks in the previous one and two years, and the total number of first-round picks in the previous three years. Lastly, measures such as luxury tax, salary cap, and preseason predicted points are also included in the data set to determine whether they influence the chances of teams in making the playoffs.

There were 30 teams in the NHL for years of 2014-2017, and 31 teams for years of 2018-2019 as the Vegas Golden Knights joined the NHL as an expansion team in 2018. Whenever possible, we will analyze the data year by year so that any changes in the pattern can be seen over time. To build various predictive models for the regular season, however, data collected in one year (30 or 31 records) may not be sufficient to come up with meaningful results. Hence, we will combine 6 years' data to have 182 records (or instances) to conduct the analysis.

The list of 32 attributes/variables in our NHL data set is as follows:

- Names of NHL teams
- Preseason predicted total points for teams
- Number of wins in the regular season
- Number of overtime losses in the regular season
- Number of total points acquired in the regular season
- Average age of players in a team

- Logarithm of the salary of a team's goaltenders (Logarithm is to the base e)
- Logarithm of the salary of a team's defensemen
- Logarithm of the salary of a team's forwards
- Logarithm of the salary of a team's centers
- Logarithm of the salary of a team's right wings
- Logarithm of the salary of a team's left wings
- Logarithm of team payroll (Goaltenders, Defensemen, and Forwards)
- Logarithm of team salary cap
- Luxury tax, in millions, paid by teams
- Indicator variable with 1 (All-In) and 0 otherwise
- Indicator variable with 1 (Believers) and 0 otherwise
- Indicator variable with 1 (One-Foot-In) and 0 otherwise
- Indicator variable with 1 (Skeptics) and 0 otherwise
- Index of sports analytics: 4 (All-In), 3 (Believers), 2 (One-Foot-In), 1 (Skeptics), 0 (Non-Believers)
- Number of analytics staff hired by a team
- Number of professional staff hired by a team
- Grade of a team's goalies evaluated before the regular season
- Grade of a team's defensemen evaluated before the regular season
- Grade of a team's forwards evaluated before the regular season
- Number of first-round picks by teams before the current season
- Highest pick in the draft by teams before the current season
- Number of first-round picks signed by teams before the current season
- Number of first-round picks by teams one year before
- Number of first-round picks by teams two years before
- Total number of first-round picks in the previous three years
- Indicator variable with 1 (playoffs) and 0 (no playoffs)

EFFECTS OF SPORTS ANALYTICS IN THE REGULAR SEASON

Conditional probabilities

Analytics belief

As mentioned before, The Great Analytics Rankings classified NHL teams' analytics belief in five categories: All-In, Believers, One-Foot-In, Skeptics, and Non-Believers. To simplify the study of conditional probabilities in determining the relationship between sports analytics belief and success of teams in the regular season, we place the categories of sports analytics belief into two groups: BELIEVERS (All-In, Believers) and NON-BELIEVERS (One-Foot-In, Skeptics, Non-Believers). There are 14 teams identified as BELIEVERS (denoted by B) and the remaining 16 teams as NON-BELIEVERS (denoted by NB). The Vegas Golden Knights have no analytics ranking classified as the expansion team joined the NHL in 2018.

The distribution of B and NB teams in no playoffs and playoffs for 2014-2019 is shown in Fig. 1. Based on Fig.1, conditional probabilities can be estimated in Table 1 to show the chances of getting into playoffs when teams are B or NB. For example, in 2014, the conditional probability of getting into playoffs for teams identified as B is P(Playoffs|B) = 8/(6 + 8) ≈ 57%. From Table 1, we see that the chances of advancing to the playoffs for a B team were higher than those for a NB team.

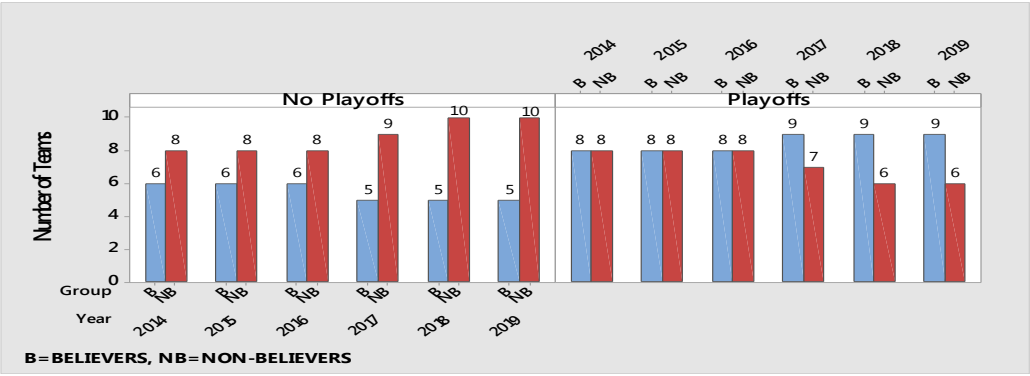

FIGURE 1. *Distribution of BELIEVERS and NON-BELIEVERS teams in no playoffs and playoffs for 2014-2019.*

Table 1. *Conditional probabilities for advancing to playoffs on the condition that teams are BELIEVERS (B) or NON-BELIEVERS (NB) for 2014-2019*

Conditional Probability	2014	2015	2016	2017	2018	2019
P(Playoffs|B)	57%	57%	57%	64%	64%	64%
P(Playoffs|NB)	50%	50%	50%	44%	38%	38%

Table 2 displays the results of testing the independence between categories of analytics belief (B or NB) and advancement to the postseasons of 2014-2019. It provides the observed values of the chi-square test statistic and the p-values. There is insufficient evidence, at the significance level of 5%, to reject the null hypothesis of independence between categories of analytics belief and advancement to the postseason. Hence there is insufficient evidence to support that advancement to the postseason was related to the categories of analytics belief of NHL teams.

Table 2. *Chi-square test for independence between categories of analytics belief and advancement to the postseason*

Year	Chi-square Statistic	P-value
2014	0.153	0.696
2015	0.153	0.696
2016	0.153	0.696
2017	1.265	0.261
2018	2.143	0.143
2019	2.143	0.143

Analytics staff

According to the information listed in the Official NHL Record Book, we obtain the number of analytics staff hired by each team for 2014-2019. Fig. 2 shows the distribution of teams having different numbers of analytics staff (0, 1, 2 or more) for advancing to playoffs or not. It is evident from the figure that teams having no analytics staff were in higher numbers compared to teams having 1 or at least 2 analytics staff in both playoffs and no playoffs for these 6 years. Also, for the years of 2014-2016, there were a lot of teams (29, 24, 22) without any analytics personnel. But for the years of 2017-2019, it appears that some of the teams without any analytics personnel in previous years decided to start hiring analytics staff to join their research and development departments. On the other hand, there were not many teams having at least 2 analytics staff. The numbers of teams having at least 2 analytics staff were 0, 1, 1, 1, 1, and 2, respectively, for 2014-2019. This might suggest that most teams did not regard hiring at least 2 analytics staff as an important contributing factor to teams' success in the regular season. The conditional probabilities in Table 3 show the chances of getting into playoffs for teams having different numbers of analytics staff (A=0,1,≥2) for 2014-2019.

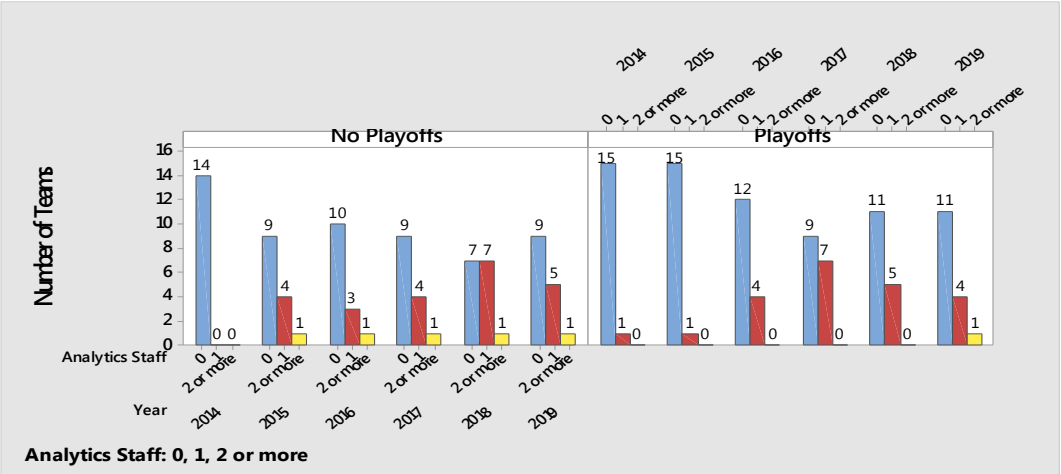

FIGURE 2. *Distribution of teams having analytics staff: 0, 1, 2 or more, in no playoffs and playoffs for 2014-2019.*

Table 3. *Conditional probabilities for advancing to playoffs on the condition that the number of analytics staff (A) hired by teams is 0, 1, or ≥ 2 for 2014-2019, where NA=Not Applicable*

Conditional Probability	2014	2015	2016	2017	2018	2019
P(Playoffs|A=0)	52%	63%	55%	50%	61%	55%
P(Playoffs|A=1)	100%	20%	57%	64%	61%	44%
P(Playoffs|A≥2)	NA	0%	0%	0%	0%	50%

The outcome NA (Not Applicable) would be assigned to the conditional probability if no teams satisfy the condition. No teams advancing to playoffs had at least 2 analytics staff for the years of 2014-2018. In 2019, Washington Capitals having 2 analytics staff made the playoffs. Besides the performance of Washington Capitals in 2018, teams having analytics staff indicated no higher percentages of success in advancing to the playoffs.

Professional staff

Using the Official NHL Record Book, we compile the list of professional staff employed by teams such as sports psychologists, nutritionists, and speed skating coaches. The professional staff, including the analytics staff, will be considered to see the combined effect of the staff on teams' success in the regular season. Teams having different numbers of professional staff (P) are categorized into three groups: High (>4), Medium (3-4), Low (0-2). The distribution of teams for their number of professional staff can be seen in Fig. 3. It is noted that there were only two teams, Nashville Predators (2014) and Ottawa Senators (2015), having no professional staff employed. Thus, they are included in the Low (0-2) group. It is interesting to see that more professional staff were hired starting from 2017. Teams began to realize the importance of professional staff providing extra advantages and support to the players. In 2019, no teams were in the Low (0-2) group; there were 10 teams and 21 teams having 3-4 and at least 5 professional staff, respectively.

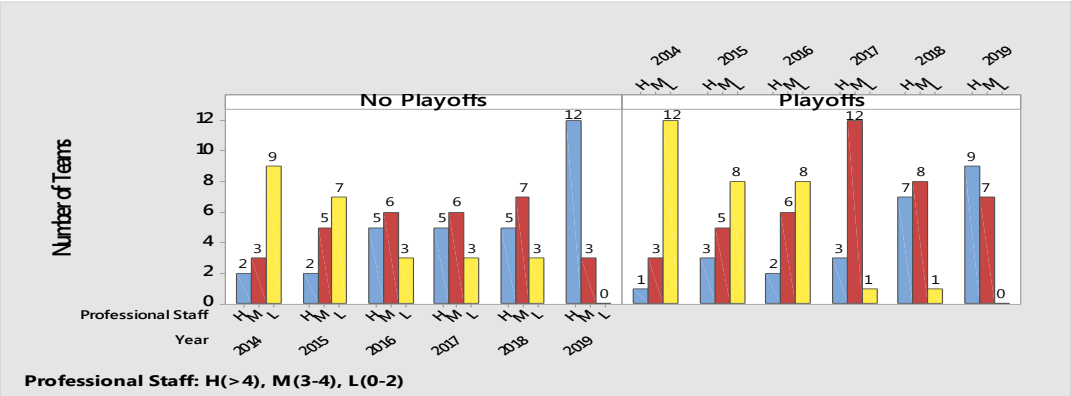

FIGURE 3. *Distribution of teams having professional staff: High(>4), Medium(3-4), Low(0-2), in no playoffs and playoffs for 2014-2019.*

The conditional probabilities in Table 4 show the chances of getting into playoffs for teams having different numbers of professional staff (P=High, Medium, Low) for 2014-2019.

Table 4. *Conditional probabilities for advancing to playoffs on the condition that the number of professional staff (P) hired by teams is High (>4), Medium (3-4), or Low (0-2) for 2014-2019, where NA = Not Applicable*

Conditional Probability	2014	2015	2016	2017	2018	2019
P(Playoffs|P=High)	33%	60%	29%	38%	58%	43%
P(Playoffs|P=Medium)	50%	50%	50%	67%	53%	70%
P(Playoffs|P=Low)	58%	53%	73%	25%	25%	NA

From Table 4, it is inconclusive to show whether the number of professional staff employed influences the chances of teams advancing to playoffs. There were 2 years for each group of professional staff to have the highest percentage of advancing to playoffs:

High (2015, 2018), Medium (2017, 2019), and Low (2014, 2016). Teams with the most professional staff hired (>4) did not necessarily achieve better success in the regular season.

CORRELATIONS

To study the Pearson sample correlations between different variables of NHL teams, we define the following notations:

N-Number of total points acquired in the regular season, i.e., 2x number of regular-season wins + 1x number of overtime losses;

T-Team payroll;

B-Categories of analytics belief: 4 (All-In), 3 (Believers), 2 (One-Foot-In), 1 (Skeptics), and 0 (Non-Believers);

A-Number of analytics staff hired;

P-Number of professional staff (including analytics staff) hired.

The Pearson sample correlation coefficients for various pairs of variables for 2014-2019 are computed and displayed in Table 5. For example, in 2014, r(N and T) = 0.74 (0.00) means that the Pearson sample correlation coefficient for assessing the linear relationship between N (number of total points acquired in the regular season) and T (team payroll) is 0.74 with p-value = 0.00. Therefore, it implies that the correlation between N and T is positive and significant at 5%. Here we are mostly interested in those pairs of variables whose correlation coefficients are significant at 5%.

Table 5. *Pearson sample correlation coefficients with p-value in parentheses for various pairs of variables for 2014-2019, where N=number of total points, T=team payroll, B=categories of analytics belief, A=number of analytics staff, and P=number of professional staff*

Relation	2014	2015	2016	2017	2018	2019
N and T	0.74(0.00)	0.66(0.00)	0.54(0.00)	0.44(0.02)	0.68(0.00)	0.70(0.00)
N and B	-0.13(0.51)	-0.10(0.60)	0.08(0.67)	0.47(0.01)	0.18(0.35)	0.31(0.09)
N and A	0.18(0.33)	-0.16(0.41)	0.003(0.99)	-0.08(0.67)	-0.22(0.24)	0.06(0.73)
N and P	-0.22(0.23)	-0.11(0.56)	-0.18(0.33)	0.03(0.86)	-0.11(0.57)	-0.21(0.26)
T and B	0.12(0.52)	0.06(0.74)	0.20(0.30)	0.05(0.79)	0.32(0.08)	0.49(0.01)
T and A	0.18(0.35)	-0.16(0.41)	-0.07(0.70)	0.21(0.26)	-0.21(0.25)	0.14(0.44)
T and P	-0.24(0.21)	0.13(0.49)	-0.13(0.49)	0.05(0.78)	-0.03(0.87)	-0.02(0.93)
B and A	0.37(0.04)	-0.19(0.32)	-0.15(0.42)	-0.02(0.90)	-0.12(0.53)	0.68(0.00)
B and P	0.34(0.06)	0.46(0.01)	0.33(0.08)	0.36(0.05)	0.28(0.14)	0.28(0.13)
A and P	0.40(0.03)	-0.05(0.81)	0.15(0.43)	0.20(0.28)	0.05(0.78)	0.42(0.02)

It is interesting to note that the team payroll (T) and N were significantly positively correlated for the years of 2014-2019 with r = 0.74, 0.66, 0.54, 0.44, 0.68, and 0.70, respectively. It suggests that the team payroll has a positive impact on the number of total points scored in the regular season. Categories of analytics belief (B) and the number of professional staff (P) indicate significant positive correlations with r = 0.46, 0.36 for 2015 and 2017. They also reveal almost significant positive correlations with r=0.34, 0.33, 0.28, 0.28 for the years of 2014, 2016, 2018-2019, respectively. This implies that the categories of analytics belief have a positive linear association with the number of professional staff employed by teams. This might suggest that the higher the category of analytics belief of a team, the greater the number of professional staff would be employed by the team. N and B were significantly correlated for 2017 (r = 0.47) but were not significantly correlated for other years. The number of analytics staff (A) and P display some significant positive correlations (r = 0.40, 0.42) for 2014 and 2019, but not the other years.

Binary logistic regression models

We employ binary logistic regression models to assess the relationship between the success of advancing to playoffs and the use of sports analytics (categories of analytics belief, number of analytics staff, and number of professional staff) for the data of 2014-2019. The Vegas Golden Knights will not be included in this study for 2018-2019 as this team has no analytics ranking. The response variable is the advancement to playoffs, which is 1 if a team advances to playoffs and 0 otherwise. The continuous explanatory variable X_1 is the categories of analytics belief, having values: 4 (All-In), 3 (Believers), 2 (One-Foot-In), 1 (Skeptics), 0 (Non-Believers). The other

continuous explanatory variables X_2 and X_3 represent the number of analytics staff and the number of professional staff, respectively. The equation of the binary logistic regression model is

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,$$

Where log is the natural logarithm; π is the probability that a team advances to playoffs; $\pi/(1-\pi)$ is the odds; β_i, i = 0, 1, 2, 3, are regression parameters to be estimated. The parameter estimates with their standard error in parentheses, odds ratios and their 95% confidence interval, and p-values of the models for years of 2015-2019 are displayed in Table 6. Note that the year 2014 is not included in this study because there was only one team (Chicago Blackhawks) with analytics staff employed in that year. It creates the problem of quasi complete separation on logistic regression.

Table 6. *Estimated values of β_0, β_1, β_2, β_3 with their standard error in parentheses, odds ratios (e^{β_1},e^{β_2},e^{β_3}) and their 95% confidence interval (CI), and p-values of the binary logistic regression models for 2015-2019*

Year	β_0	β_1	β_2	β_3	e^{β_1}	e^{β_2}	e^{β_3}	P-value
2015	0.67	-0.04	-2.01	-0.02	0.96	0.13	0.99	0.20
	(1.23)	(0.54)	(1.16)	(0.23)	(0.34, 2.74)	(0.01, 1.30)	(0.63, 1.54)	
2016	-0.19	0.55	-0.12	-0.26	1.72	0.89	0.77	0.50
	(1.25)	(0.53)	(0.74)	(0.21)	(0.61, 4.84)	(0.21, 3.80)	(0.51, 1.17)	
2017	-0.74	0.25	-0.03	0.07	1.28	0.97	1.07	0.86
	(1.18)	(0.49)	(0.69)	(0.16)	(0.50, 3.32)	(0.25, 3.71)	(0.79, 1.46)	
2018	-0.01	0.11	-0.89	0.03	1.12	0.41	1.03	0.60
	(1.23)	(0.47)	(0.71)	(0.13)	(0.44, 2.83)	(0.10, 1.65)	(0.80, 1.33)	
2019	-0.38	0.58	-0.06	-0.14	1.78	0.95	0.87	0.52
	(1.29)	(0.50)	(0.69)	(0.14)	(0.67, 4.76)	(0.25, 3.63)	(0.66, 1.14)	

We test the null hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$, and all p-values are not significant at 5% for 2015-2019. Hence there is insufficient evidence to conclude that at least one of the categories of analytics belief, the number of analytics staff, and the number of professional staff employed was associated with a team's success in advancing to playoffs. It seems that these measures of sports analytics do not have much impact on the success of a team in the regular season.

The associated goodness-of-fit tests are shown in Table 7. For Pearson and Hosmer-Lemeshow tests, there is insufficient evidence (at 5% level of significance) to claim that the binary logistic regression models do not fit the data adequately for 2015-2019. However, the p-values of the Deviance test are very close to 0.05 for 2015-2019. These three tests give mixed signals on whether the binary logistic regression models fit the data adequately for these years.

Table 7. *P-values of the associated goodness-of-fit tests for binary logistic regression models for 2015-2019*

Year	Pearson	Deviance	Hosmer-Lemeshow
2015	0.30	0.08	0.83
2016	0.26	0.05	0.09
2017	0.28	0.03	0.80
2018	0.28	0.04	0.52
2019	0.22	0.05	0.14

Predictive Modeling In The Regular Season

Decision trees

All team attributes available

The target variable is the advancement to playoffs, which has class labels: 1 if a team advances to playoffs and 0 otherwise. We will include all the input variables, listed in Section 2, to train a classification model with 182 instances for the cumulative 6 years' data of the regular season. SAS Enterprise Miner Workstation 14.2 is used to implement the algorithm of a decision tree to obtain the optimal decision tree, without overfitting the classification model, for classifying teams into playoffs or no playoffs. The resulting decision tree is displayed in Fig. 4.

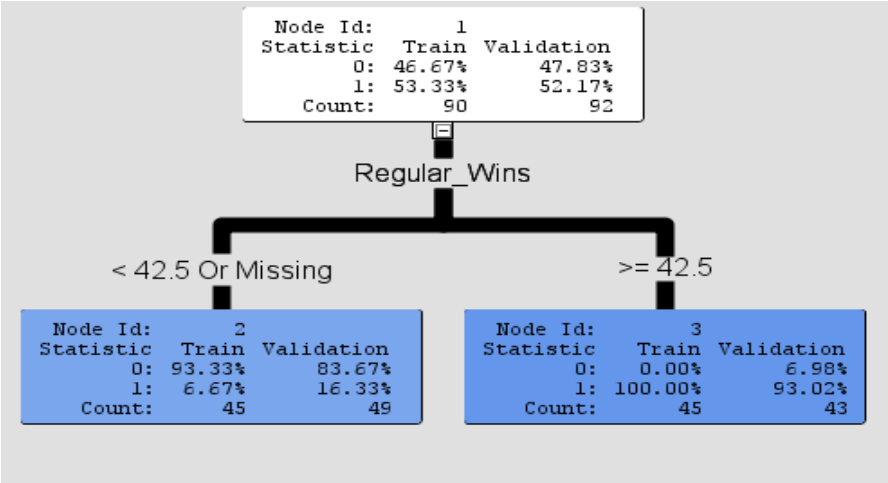

FIGURE 4. *Decision tree based on all team attributes available.*

Node 1 shows that the data are randomly partitioned into the training model (90 teams) and the validation model (92 teams). Among the 90 teams in Train, 46.67% of them were not in playoffs while 53.33% of them were. For the 92 teams in Validation, 47.83% of them were not in playoffs while 52.17% of them were. It shows that the numbers of records for Train and Validation in each class label are roughly the same. We are more concerned with the validation error rate since it reflects a more general and realistic situation for classifying teams into playoffs or not based on the chosen attributes and split conditions.

The decision tree indicates that the number of regular-season wins is first chosen among all those input variables to split the tree. Since there are no missing observations in our data set, the corresponding split conditions are < 42.5 (i.e., less than or equal to 42 wins in a regular season) and ≥ 42.5 (i.e., greater than or equal to 43 wins in a regular season). The 90 teams in Train are separated into 45 teams in Node 2 and 45 teams in Node 3. Likewise, the 92 teams in Validation are separated into 49 teams in Node 2 and 43 teams in Node 3. In Node 2 (wins < 42.5), 6.67% of the 45 teams in Train and 16.33% of the 49 teams in Validation advanced to playoffs. In Node 3 (wins ≥ 42.5), all 45 teams in Train and 93.02% of the 43 teams in Validation advanced to playoffs. Therefore, we may conclude that teams winning 43 games or more in a regular season are very likely to move forward to playoffs, while teams winning 42 games or less in a regular season are not likely to move forward to playoffs.

Applying this decision tree to our data set, SAS produces the misclassification error rates for the training and validation models as 3.33% and 11.96%, respectively. As a result, we conclude that the test conditions (< 42.5 and ≥ 42.5) of the regular-season wins are very good criteria to classify teams into playoffs or no playoffs.

All team attributes excluding the number of regular-season wins

The number of regular-season wins is not the determining factor for choosing teams into playoffs or not. Rather, it is the number of total points acquired (2 x number of regular-season wins + 1 x number of overtime losses) that mostly determines the top 16 teams advancing to the playoffs. The interactive mode of SAS Enterprise Miner is then applied to train the decision tree to classify teams into playoffs or not, based on all those team attributes used in the previous section but excluding the regular-season wins. The result is given in Fig. 5.

Node 1 displays the same partition for Train and Validation as in Fig. 4. The decision tree in Fig. 5 shows that the number of total points acquired is first chosen to split the tree. The corresponding test conditions are < 95.5 (i.e., less than or equal to 95 points acquired) and ≥ 95.5 (i.e., greater than or equal to 96 points acquired). The 90 teams in Train are separated into 45 teams in Node 2 and 45 teams in Node 3. Likewise, the 92 teams in Validation are separated into 52 teams in Node 2 and 40 teams in Node 3. In Node 2 (points < 95.5), 6.67% of the 45 teams in Train and 21.15% of the 52 teams in Validation advanced to playoffs. In Node 3 (points ≥ 95.5), all 45 teams in Train and 92.50% of the 40 teams in Validation advanced to playoffs.

The algorithm then chooses Points again to

further split Node 2 with the test conditions: < 92.5 and ≥ 92.5. For those teams with less than 92.5 points in Node 7, none of the 38 teams in Train and 6.98% of the 43 teams in Validation advanced to playoffs. For those teams with at least 92.5 points in Node 8, 42.86% of the 7 teams in Train and 88.89% of the 9 teams in Validation advanced to playoffs.

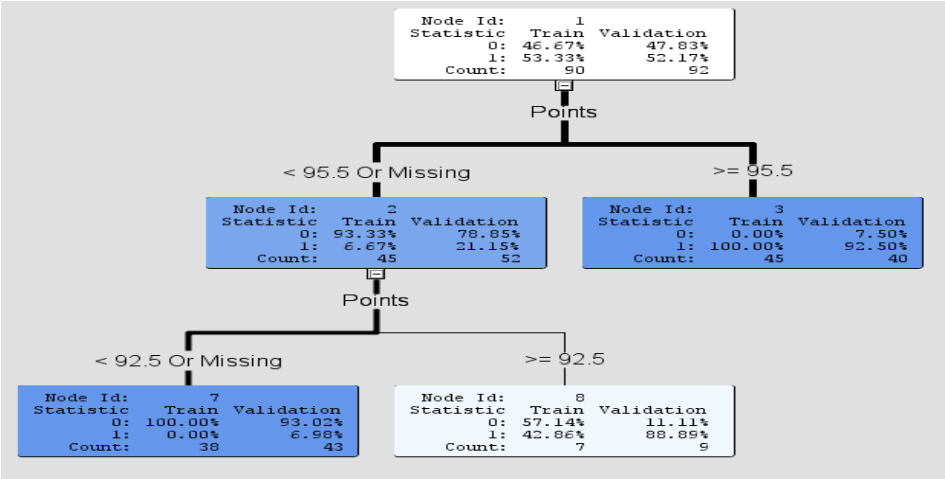

FIGURE 5. *Decision tree based on all team attributes excluding the number of regular-season wins.*

Therefore, we may conclude that teams having 96 points or more in a regular season are very likely to advance to playoffs, while teams having points between 93 and 95 points (inclusive) can still have a decent to pretty good chance (42.86% in Train to 88.89% in Validation) to make the playoffs, depending on the distribution of points and the configuration of wild card teams. Teams having 92 points or less in a regular season are very unlikely to advance to playoffs. The misclassification error rates for Train and Validation are 3.33% and 15.22%, respectively. Consequently, we conclude that the test conditions (< 92.5, ≥ 92.5 but < 95.5, and ≥ 95.5) of the total points acquired are very good criteria to classify teams into playoffs or no playoffs.

When regular-season wins and total points not available at the beginning of a season

At the beginning of an NHL season in mid-October, the information on the number of regular-season wins and total points acquired by teams is certainly not available. In this situation, one may still wish to build a decision tree to classify teams into playoffs or no playoffs. SAS produces the optimal decision tree based on those input variables described in the previous section excluding the number of regular-season wins and total points. The result is given in Fig. 6.

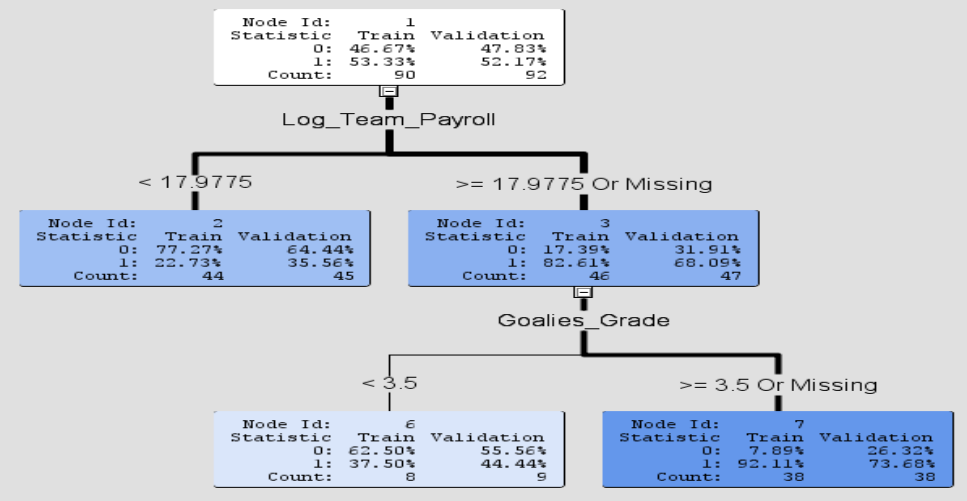

FIGURE 6. *Decision tree when regular-season wins and total points not available at the beginning of a season.*

Node 1 displays that data are randomly divided into Train (90 teams) and Validation (92 teams). The first input variable chosen to split the tree is Log Team Payroll. The test conditions are <

17.9775 and ≥ 17.9775 that is equivalent to $64,199,116 (≈$64.2m) for team payroll. Under these test conditions, the 90 teams in Train are separated into 44 teams in Node 2 and 46 teams in Node 3. Similarly, the 92 teams in Validation are separated into 45 teams in Node 2 and 47 teams in Node 3. In Node 2 (team payroll < $64.2m), 22.73% of the 44 teams in Train and 35.56% of the 45 teams in Validation advanced to playoffs. In Node 3 (team payroll ≥ $ 64.2m) ,however, 82.61% of the 46 teams in Train and 68.09% of the 47 teams in Validation advanced to playoffs. Consequently, the team payroll with a threshold of $64.2m played an important role to identify teams whether they would have higher chances of moving forward to the playoffs.

The decision tree algorithm in SAS further splits Node 3, and the input variable chosen is Goalies Grade with test conditions: < 3.5 and ≥ 3.5. The expected performances of teams' goalies (such as save percentages, goals against average, and shutouts) were estimated and given in terms of letter grades by The Hockey News Magazine before the start of a regular season. To facilitate the decision tree algorithm, the letter grades are converted to the equivalent numerical values shown in Table 8. Note that the value 3.5 falls between grades C and C+.

Table 8. *The Hockey News Magazine's letter grades for Goalies, Defensemen, and Forwards and their numerical equivalency*

Letter Grade	Numerical Equivalency
A+	10
A	9
A-	8
B+	7
B	6
B-	5
C+	4
C	3
C-	2
D	1

Under the test conditions for splitting Goalies Grade, the 46 teams in Train are separated into 8 teams in Node 6 and 38 teams in Node 7. Likewise, the 47 teams in Validation are separated into 9 teams in Node 6 and 38 teams in Node 7. For teams with Goalies Grade < 3.5, Node 6 shows that 37.50% of the 8 teams in Train and 44.44% of the 9 teams in Validation advanced to playoffs. However, Node 7 (Goalies Grade ≥ 3.5) shows that 92.11% of the 38 teams in Train and 73.68% of the 38 teams in Validation advanced to playoffs. This reveals that in addition to a team's spending of at least $64.2m on players' salaries, acquiring decent goalies with a grade of at least 3.5 (C+ or above) would significantly increase the team's chances of advancing to playoffs.

The misclassification error rates for the training and validation models are 17.78% and 32.61%, respectively. This decision tree results in moderately accurate predictions for classifying teams into playoffs or no playoffs when regular-season wins and total points are not available at the beginning of a season.

It is more interesting to examine the effects of all those input variables available at the beginning of an NHL season. Thus, the number of regular-season wins and total points will not be considered in the sections below.

The random forest

The random forest functionality (HP Forest) of SAS Enterprise Miner is employed to train multiple decision trees collectively to classify teams into playoffs or no playoffs, based on those input variables described in the previous sections. There are 29 input variables involved in a random forest. The target variable is the advancement to the playoffs.

The HP Forest employs a procedure that selects 5 input variables at a time (by default) from the collection of the 29 input variables. This procedure also uses an in bag fraction of 60% (by default). The in bag fraction refers to the percentage of bagging procedure, which selects 60% of the Train data with replacement, to ensure that decision trees are trained independently and efficiently by the Train data. A collection of 2,000 decision trees is built to classify teams into playoffs or no playoffs, and a majority vote system is used to determine the class of a particular record. In this case, a

building 2,000 decision tree is required to make sure that the Validation misclassification error rate converges.

The misclassification error rates for Train and Validation of 2,000 decision trees are 0% and 30.43%, respectively. Therefore, the Validation misclassification error rate of the random forest is lower than the corresponding error rate (32.61%) of the single decision tree developed in Section 4.1.3. In this situation, building a random forest of 2,000 decision trees does improve the misclassification error rate.

The logistic regression with stepwise selection

We utilize a binary logistic regression model to assess the relationship between a team's success in advancing to playoffs and those 29 input (or explanatory) variables described in the previous section. Running a logistic regression model with these 29 variables will lead to model overfitting. As a result, the method of stepwise selection is used to choose the most significant variables for the regression model.

After running the stepwise selection in the logistic regression model in SAS, only two significant explanatory variables are chosen to form the linear relationship with the log odds of making the playoffs. These two variables are Log Team Payroll (X_1) and Goalies Grade (X_2). The corresponding equation of the binary logistic regression model is

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

where π is the probability that a team advances to playoffs. Table 9 presents the result of fitting the regression model for the cumulative regular-season data of 2014-2019. Based on the p-value, there is sufficient evidence, at 5% level of significance, to conclude that Log Team Payroll and Goalies Grade as a whole is associated with the success of teams advancing to playoffs in the binary logistic regression model. This suggests that before the start of a season, the team payroll and the grade of a team's goalies as a whole is a good contributing factor to whether the team makes the playoffs or not. The outcome of selecting these two variables in the logistic regression model is the same as the result obtained from the decision tree in the previous section.

Table 9. *Estimated values of β_0, β_1, β_2 with their standard error in parentheses, odds ratios (e^{β_1},e^{β_2}) and their 95% CI, and p-value of the binary logistic regression model with stepwise selection for the cumulative regular-season data of 2014-2019*

β_0	β_1	β_2	$e^{\beta 1}$	$e^{\beta 2}$	P-value
-311.50	39.60	0.56	1.6E17	1.75	<0.0001
(70.85)	(9.04)	(0.17)	(3.2E9, 7.8E24)	(1.25, 2.45)	

The misclassification error rates for the training and validation models are 18.89% and 32.61%, respectively. Consequently, the logistic regression model with Log Team Payroll and Goalies Grade as explanatory variables is a moderately accurate model to predict whether a team will advance to playoffs or not.

The neural network with stepwise selection

An artificial neural network, a non-linear regression model, will also be considered for the classification problem. The neural network with the TANH activation function from SAS Enterprise Miner is applied to build a neural network via the training records. This neural network is then used to predict teams, based on the team attributes, whether they make the playoffs or not for the validation records.

Instead of running a neural network with 29 input variables, the method of stepwise selection is used to choose the variable(s) that exhibit a significant relationship with the target variable (playoffs or no playoffs). Log Team Payroll (X_1) and Goalies Grade (X_2) are the only two explanatory variables chosen. Then we employ Auto Neural in SAS to determine how many neurons to be used in a single hidden layer. In this case, three neurons are needed. The Neural Network in SAS is then applied to determine the associated weights of the chosen variables. The estimated neural network equation and the activation equations are

$$\log\left(\frac{\pi}{1-\pi}\right) = \widehat{W_0} + \widehat{W_1}H_1 + \widehat{W_2}H_2 + \widehat{W_3}H_3$$

$$H_1 = \tanh(\widehat{W_{10}} + \widehat{W_{11}}X_1 + \widehat{W_{12}}X_2),$$

$$H_2 = \tanh(\widehat{W_{20}} + \widehat{W_{21}}X_1 + \widehat{W_{22}}X_2),$$

$$H_3 = \tanh(\widehat{W_{30}} + \widehat{W_{31}}X_1 + \widehat{W_{32}}X_2).$$

The estimated weights are as follows:

$$\widehat{W_0} = 6.94,$$
$$\widehat{W_1} = -11.79,$$
$$\widehat{W_2} = -15.24,$$
$$\widehat{W_3} = 15.87,$$
$$\widehat{W_{10}} = -13.33,$$
$$\widehat{W_{11}} = 9.83$$
$$\widehat{W_{12}} = -20.73,$$
$$\widehat{W_{20}} = 0.33,$$
$$\widehat{W_{21}} = -0.20,$$
$$\widehat{W_{22}} = -0.23,$$
$$\widehat{W_{30}} = -1.28,$$
$$\widehat{W_{31}} = -0.18,$$
$$\text{and } \widehat{W_{32}} = -1.53.$$

The misclassification error rates for the training and validation models are 14.44% and 30.43%, respectively. This neural network has a lower misclassification error rate than that of the corresponding logistic regression model. However, both the random forest and the neural network with stepwise selection generate the same misclassification error rate for the validation records.

Conclusion And Comments

The categories of analytics belief and the number of analytics staff hired haven't shown any significant relationship with the success of teams in advancing to playoffs. It appears from Fig. 3 that more professional staff (including the analytics staff) has been hired by teams since 2017. That seeks to improve teams' performance and strengthen their chances of success. It looks promising for teams with the medium number (3-4) of professional staff as they performed more consistently for teams' success in terms of advancing to playoffs as compared to the other two groups. This observation needs to be validated by more data. The regular-season wins and total points scored have been identified as essential predictors of a team's success in the regular season, in terms of advancing to playoffs. It makes sense as these two attributes provide useful quantitative information on how well (or poorly) a team performs in the regular season. They form a strong basis for determining whether a team makes the playoffs or not. But in reality, the total points scored would mostly be the determining factor for advancing to the postseason. Based on the decision tree in Fig. 5, teams having 96 points or more are very likely (92.50%) advancing to playoffs, while teams having points in the range of 93-95 still have a pretty good chance to make the playoffs, depending on the distribution of points and the configuration of wild card teams in that particular season. Teams having 92 points or less are very unlikely (6.98%) advancing to playoffs. Therefore, the test conditions (< 92.5 and ≥ 95.5) of the input variable Points are excellent criteria to classify NHL teams into playoffs or no playoffs. These benchmarks could serve as points of reference for teams to gauge their likelihood of making the playoffs.

When the regular-season wins and total points scored are not available at the beginning of a season, the team payroll (after logarithm) and Goalies Grade have been identified as important factors for a team's success in the regular season. The decision tree in Fig. 6 reveals that teams with a team payroll of at least \$64.2m would have 82.61% and 68.09% advancing to playoffs in Train and Validation, respectively. In addition to this important factor, teams with Goalies Grade of 3.5 or higher (at least a C+ in letter grades defined by The Hockey News Magazine) significantly improve their chances of advancement to 92.11% and 73.68% in Train and Validation, respectively.

Random forests appear to provide lower validation misclassification error rates for classifying teams into playoffs or not than does a single decision tree. A random forest might need to generate a large number of decision trees for the validation misclassification error rate to converge.

This number of trees generated, however, might vary depending upon the situation. The neural network with stepwise selection, as anticipated, is better than or as good as the binary logistic regression with stepwise selection with respect to the validation misclassification error rate.

Using hidden neurons in a single hidden layer most likely improves the capability of a predictive model to describe the complicated relationship between the explanatory and target variables. However, there is no easy way to interpret the significance of the weights of variables in a neural network.

Random forests appear to be the best or as good

as the other three predictive modeling techniques to have the lowest validation misclassification error rate. Finally, the predictive modeling techniques considered in this study require a large data set to better train the models. Our current NHL data set, however, consists of only six years' data from 2014 to 2019, with 182 team records in the regular season. Therefore, the results obtained here are preliminary and they need to be confirmed by more data in the future.

REFERENCES

[1] Chu, D., Wang, C., 2019. Empirical study on relationship between sports analytics and success in regular season and postseason in Major League Baseball. Journal of Sports Analytics, (5) 205-222.

[2] Hockey-reference.com, 2014-2019. 'NHL Standings'.URL:https://www.hockeyreferene .com/leagues/NHL_2014_standings.html(Substitute 2015-2019 for 2014 to get the corresponding URL.)

[3] Espn.com, 2015. 'The Great Analytics' Rankings'.URL:http://www.espn.com/espn/feature/story?_slug_=the-great-analytics-rankings & id= 12331388 & redirected=true

[4] Hockey-reference.com, 2014-2019. 'NHL Playoffs' URL:https://www.hockey- reference .com/playoffs/NHL_2014.html (Substitute 2015-2019 for 2014 to get the corresponding URL.)

[5] Schwartz, N.,Zarrow,J.,2009. An analysis of the impact of team payroll on regular season and postseason success in Major League Baseball. Undergraduate Economic Review, Vol.5:Iss.1, Article3. URL:http://digitalcommons. iwu.edu/uer/vol5/iss1/3

[6] Spotrac.com, 2014-2019. 'NHL Team Payroll'.URL:https://www.spotrac.com /nl/cap/2014/l (Substitute 2015-2019 for 2014 to get the corresponding URL.)

[7] The Hockey News Magazine, 2014-2019. Edition: Playoff Pictures, Prospects. Information of NHL team grades.

[8] Official NHL Record Book, 2014-2019. Analytics staff and professional staff information. URL: https://vpl.biblioc ommons.com/item/show/220863038

[9] NHL Draft Picks, 2014-2019. URL: https://www.capfriendly.com/draft/20 14 (Substitute 2015-2019 for 2014 to get the corresponding URL.)

Citation: *David Chu, Gurdeepak Sidhu. "Effects of the Use of Sports Analytics and Team Attributes on Success in Regular Season of National Hockey League", Journal of Sports and Games, 3(2), 2021, pp.1-13. DOI: https://doi.org/10.22259/ 2642-8466.0302001*