NHL Expected Goals Model

Luke Moslenko

501135769

Final Project - ES8913

April 18th, 2022

Introduction

Hockey is one of the most popular sports in North America with the NHL being among the top 4 professional league on the continent. According to the International Ice Hockey Federation (IIHF), 1.64 million people play hockey around the world. The sport has continued to rise in popularity with over 23 professional leagues around the world. Hockey is also Canada's national winter sport, embedding itself in the culture of the country. Hockey is a unique game based on a number of categories. It is one of the few team sports that does not use a ball, instead using a puck. It is also one of a few team sports that is played on ice. The players also wear more protective equipment that nearly any other sports, mostly due to the nature of a collision and the velocity of the puck when shot which can exceed 160km/h. Hockey is a team sport, with each team consisting of 5 skates and a goaltender, During the game, players use sticks to direct the puck towards the net with the goal of putting the puck into the net (Pichedda, 2014). All of the rules and specificity surrounding hockey all lead to one goal, to score into the opposing team's net more than they score on you.

While the goal of a hockey game may sound simple, the challenge with modeling and predicting hockey is a combination of randomness and low scoring rates (Liu et al. 2020). When analyzing a game, the majority of the time is spend performing actions that indirectly cause or suppress goals and scoring chance. While this is the majority of events, some of the most common statistics recorded are goals and assists, documenting a very coarse overview of the game. Additionally, hockey is a highly integrated team games where all the players on the ice regularly handle the puck and all players offer a specific contributing to the overall success of the team. There is a near infinite combinations of strategic approaches to produce, making each sequence of events nearly unique. During a single game, approximately 6 goals are scored when the total shots on goal regularly exceeding 60, with numerous games having 2 or fewer goals. The low scoring rates exacerbate the challenge of modeling over an entire season or more of data. The lack of description from traditional hockey statistics has pushed hockey analysts to explore other methods to better describe the game. Shots on goal are commonly used to evaluate a team's performance and to make predictions about future performance due to the abundance of shots during a game and the supporting situational data associated with each shot (MacDonald, 2012). Using shots allows for a better wholistic description of a player or team's performance than goals alone.

Numerous approaches have been offered to add more context to a player's individual contribution and a team's overall performance during a game. Because there is heterogeneity in the amount of time each player is on the ice during a game, a number of metrics have attempted to normalize the temporal component to a specific unit of team, typically 60 minutes. For example, goals per 60 minutes or shots per 60 minutes are calculated to determine the number of goals or shots a player would score during 60 minutes on the ice. Modeling is also used to evaluate player and team performance, with expected goals being a popular model. Expected goals calculated the probability of shot being a goal based on numerous drivers including shot

distance and shot angle (Douglas et al. 2021), with logistic regression being a common modeling technique used to evaluate NHL hockey (Gramacy et al. 2013). These probabilities are added up to calculates expected goals. There are also secondary calculations that can be made from expected goals. Goals above expected is the number of goals score that exceed the predicted value. This is a proxy measurement to estimate the skill of the player that cannot be qualified from traditional event data. Goals saved above expected is a metric used to evaluate goaltenders. This metrics uses the difference between the true goals against and the expected goals against to determine whether a goaltender allowed more or less goals than expected, indicating their level of performance based on expectations. Even more complicated models are used in an effort to quantify the entirety of the on-ice contribution of a player instead of just shots on net. For example, Dom Luszczyszyn of The Athletic developed a metric called "Game Score" which uses a weighted approached to all the events involving each number including shots on goal, body checks, blocked shots, takeaways, give aways, and penalties to determine a single number describing a player's overall performance. On a longer temporal scale, wins above replacement (WAR) is an analysis of a player's value of the course of the season (Nandakumar & Jensen, 2019). Using offensive and defensive, the number of wins a player has individually contributed to the team over the course of a season compared to a replacement level player. A replacement level player is a description of a player on the fringe of skill level to maintain in the NHL and could be easily acquired or signed in exchange for very little salary or assets.

Because the NHL is a professional league comprised of players who are paid large salaries to play hockey, a collective bargaining agreement is in place as a framework for all business dealings within the league (Idson & Kahane, 2000). A major component of the collective bargaining agreement is a maximum budget for player salaries per team, called a salary cap. For the 2020-2021 season the salary cap was set at $81.5 million. Because players negotiate contract with their current teams or on the open market, there is considerable variation in the value of contract between players. The value of a contract can be an important contributor to whether a player is beneficial to their team due a limit in spending. There are numerous cases where a player's performance drastically exceeds their contract value and cases where a player underperforms within a large salary. A key component of hockey management is evaluating the monetary value of a player with the goal of building the most successful team within the constraints of the salary cap.

In this analysis, the NHL play-by-play data will be used to create an expected goals model using logistic regression with a variety of spatial, situational, and personnel event descriptors. The goal is to quantify the contribution to offense of each individual player and to evaluate a team's performance over the course of an entire season. Using the expected goals model, the outcome of every game will be predicted for the 2020-2021 NHL season. Additionally, expected goals and player salary data will be used to evaluate the financial value of each player. Lastly,

an investigation will be conducted to identify if a relationship exists between team performance in the regular season and how a team spends money on offense.

Methods:

Data acquisition:

The data included in this study is publicly available play-by play data sourced from the NHL data repository. For this analysis, the package hockeyR (Morse, 2022) was used to scrape the play-by-play data for each season in the training and testing of the model. The hockeyR package is able to scrape play-by-play data dating back to the 2010-2011 NHL season. The raw play-by-play data from the NHL repository is event based, capturing major events that take place during the game. The recorded events included shots on goal, missed shots, blocked shots, goals, body checks, player changes, faceoffs, stoppages in play, penalties, and giveaways. A number of details are recorded for additional information and contextualization of each event. The main participants for a shot are the shooter and the goaltender, for a goal it is who scored the goal, the 1 or 2 players that assisted the goal and the goaltender, for a body check, the person delivering the check and the person receiving the check. For all events, the strength state, an indicator of the number of skaters on the ice per team, is recorded along with the time the event took place in game seconds is recorded as well as the location in x,y coordinates. To contextualized the x,y coordinates system. An NHL rink is 200 feet long and 85 feet wide. X coordinates range from -100 to 100. Indicating 1 foot per integer. The y coordinates range from –42 to 42, indicating slightly less than 1 foot per integer since half of the rink's width is 42.5ft wide.

Within the functionality of hockeyR, there are additional variables of interest added to the data to facilitate better analysis. The shot angle is calculated relative to the goal line, with a shot directly facing the net at 90 degrees and a maximum angle of 180 degrees. All values greater than 180 are corrected to 180 degrees. The shot distance is also calculated using the coordinates, indicating the distance between the location of the where the shot was taken and the location of the net, measured in feet. The final addition from the hockeyR package is an adjusted coordinates system to compensate for teams switching sides it between periods. With this adjusted system the offensive events of the home team are always located on the right side of the rink ($x > 0$) and the offensive events of the away team are always on the left ($x < 0$).

Two datasets were used in this analysis. A training dataset that used to train the model consisted of 2 seasons of NHL hockey. The 2018-2019 season and the 2019-2020 season were combined to create this data set. Once data cleaning and model filtering took place, there were 156,036 events made up of shots and goals over the course of 2365 games. The testing data set was the data used to test the model and execute predictions and analysis. This dataset comprised of the events of the 2020-2021 NHL hockey. Once data cleaning and model filtering took place, there 55,453 events made up of shots and goals over the course of 868 games.

Player Salary Data:

Player salaries were also acquired as part of the monetary value analysis of players and teams Information on player salaries were acquired from CapFriendly.com. All salary data was acquired for the 2020-2021 NHL season. Player salary is listed as two numbers, cap hit and a salary (National Hockey League, 2013). The salary is the amount of money the player receives in compensation from their team including any performance or signing bonuses. The cap hit in the amount of money that counts against the team's salary budget for that season. In most cases, these figures are the same, but discrepancies can occur due to contract structures of long-term deals where a player is paid a different amount in each year of a contract. In these cases, the cap hit is the average salary paid to the player and that value is counted against the yearly budget of the team. Because there is a maximum of the total cap hit of a team and not a direct cap on the total salary for the season, the cap hit of each player was used to assess their monetary value to their team. The result is a comprehensive dataset yearly salary data and all the major events that occur in a game with a substantial amount of supporting information to conduct a variety of studies of the game of hockey.

Data cleaning: nhldataclean

In order to clean the data and prepare the data for analysis, an in-house function called nhldataclean was built to execute a variety of data cleaning functions associates analyzing the play-by-play data and to prepare the data for modeling. The function calculated the time difference between events, allowing for specific event types to be determine and to calculate total time on ice per player and puck possession time per team. Given the event, a logical is assigned for whether the event was a goal. Since event type is given in text format, assigning goal as 1 and all other shots as 0, prepares our analysis for future logistic regression. The data is filtered for strength states that occur most commonly in the game, goaltenders are not included in the count of strength state, meaning that 5 skaters on each side is the typical orientation. Strength states were included in anomalous data that are not possible under NHL rules. The most skaters a team can have on the ice is 6 if a player replaces the goaltender, leaving the net unattended. The minimum number of skaters a team could have been 3, this occurs when team has 2 concurrent penalties. The strength states included are 6vs5, 6vs4, 6vs3, 5vs5, 5vs4, 5vs3, 4vs3, & 3vs3. For specific scenarios, the function indicates if the shot was part of a rush sequence. This function assigns a 1 for a rush and a zero for non-rush sequences. The criteria for a rush sequence are that the previous event must have occurs within the shooting team's defensive zone, with an x coordinate less than 60 and the previous event must have occurred more than 4 seconds before the current event. The function also indicated if the shot was a rebound. The functions assigned a 1 for a rebound and a zero for non-rebounds. The criteria for a rebound are that the previous event must have occurred 2 seconds or less prior to the current event, and both events must be a shot on goal. Lastly, the function assigned shot type, strength state and event goalie as a factor and removes any incomplete cases.

The data underwent a secondary filtering that was specific to the requirements of the model. Since the model is focused on shots on net to calculate expected goals, only shots and goals are needed. A subset of the previously cleaned data was taken to only included events that were shots and goals, and the remaining data was filtered to remove events without a shot type.

Model:

The model aimed to predicted expected goals based on a variety of situational, spatial, and personnel variables that could influence the quality of a shooting event. Because each shot on net has a certain probability of being a goal, the goal of the model is to estimate that probability based on the independent variables. The model should predict that a higher quality shot will have a higher probability of going in. Looking at the concept of expected goals, they are the sum of the probabilities of each shot. For example, if ten shots are taken each having a 10% probability of being a goal, it is expected that one goal will be scored. Logistic regression is the chosen technique used to model the probability of each shot taken on goal. Logistic regression is used to explain the variation of a binary variable, it is an ideal tool to model whether or not a goal was achieved and assign a probability between 0 and 1 for each event (Cox, 1958). As previously stated, the dependent variable of the model is simply whether or not the event was a goal. This was assigned a 1 for goal and a 0 for no goal in the functionality of nhldataclean. The drivers of the model included a number of potential drivers of event quality including the shot distance, shot angle, shot type, strength state, rebounds, rush chances, and the goalie that was in net for the event. For this model, shot distance and shot angle were included as a polynomial to increasingly incentivize shot that are close to the net and in front of the net while disincentivizing shots from long distances or sharp angles.

Model Evaluation:

Because logistic regression does not require the traditional assumptions of normality, homoscedasticity, or linearity found in linear regression, other methods are needed to evaluate the model (Peng et al. 2002). Residual deviance and a ROC curve were the notable methods used for model evaluation. Residual deviance in a measure of the total size of the residuals throughout the model (Kuss, 2002). A lower residual deviance is an indicator of better model fit. Using residual deviance, the model was assessed to see if a simpler model existed using a combination of existing drivers that could better explain the data. A systematic approach of remove each of the independent variables from the model one at a time and assessed the residual deviance. The result was that the model in its current form had the lowest residual deviance of all the model iterations.

Table 1: Residual deviance assessment of the expected goals logistic model

```
Start:  AIC=86555.16
is_goal ~ poly(shot_distance, 3, raw = TRUE) + poly(shot_angle,
    3, raw = TRUE) + secondary_type + strength_state + is_rebound +
    is_rush + event_goalie_name

                                      Df Deviance   AIC
<none>                                     86507 86555
- is_rush                               1    86512 86558
- event_goalie_name                     1    86551 86597
- is_rebound                            1    86671 86717
- secondary_type                        6    87018 87054
- poly(shot_angle, 3, raw = TRUE)       3    87357 87399
- strength_state                        8    87681 87713
- poly(shot_distance, 3, raw = TRUE)    3    92071 92113
```

The model was also assessed a receiver operation characteristic curve and an evaluation of the area of the curve (Bradley, 1997). The ROC curve visualized the ratio of true positives to false positives in the model at a variety of classification thresholds. A linear line appears in the plot indicating the location where the number of true positives is equal to the number of false positives. If the curve is above and to the left of the line, this indicates that the number of true positives is greater that the number of false positives. If the curve is below and the right of the line, this indicates that the number of true positives is lower than the number of false positives. The area under the curve is a calculated probability that a random positive value ranks more highly in the classification scheme that a negative value. In this case an AUC at zero indicates that all the model classifications are incorrect where an AUC of 1 indicates that the all the model classifications are correct. In this model, the ROC curve is on the left of the line for all of the classification thresholds indicating a greater number of true positives than false positives throughout the model. The model also achieved an AUC of 0.74, indicating that our classifications were correct nearly three out of 4 times.

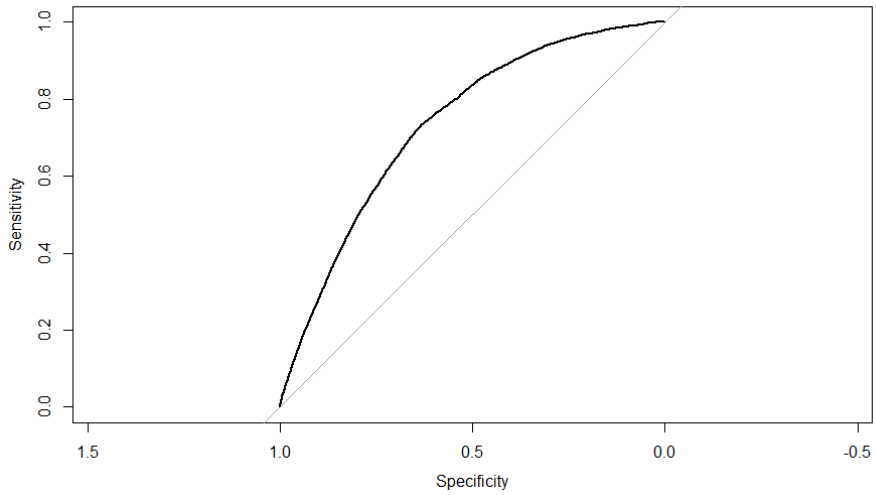

Figure 1: ROC curve depicting the ratio of true positives to false positives in the model

Results and Discussion:

For the 2020-2021 NHL the expected goals for each player were calculated using all the shots taken over the course of the season. The result was compared with the recorded goal totals for the season to a metric of model performance. Of the 896 players that played in the 2020-2021 season, the model prediction was within 2 goals of the true total for 626 players and within 1 goal for 412 players, indicate a 70% and 42% accuracy rating respectively.

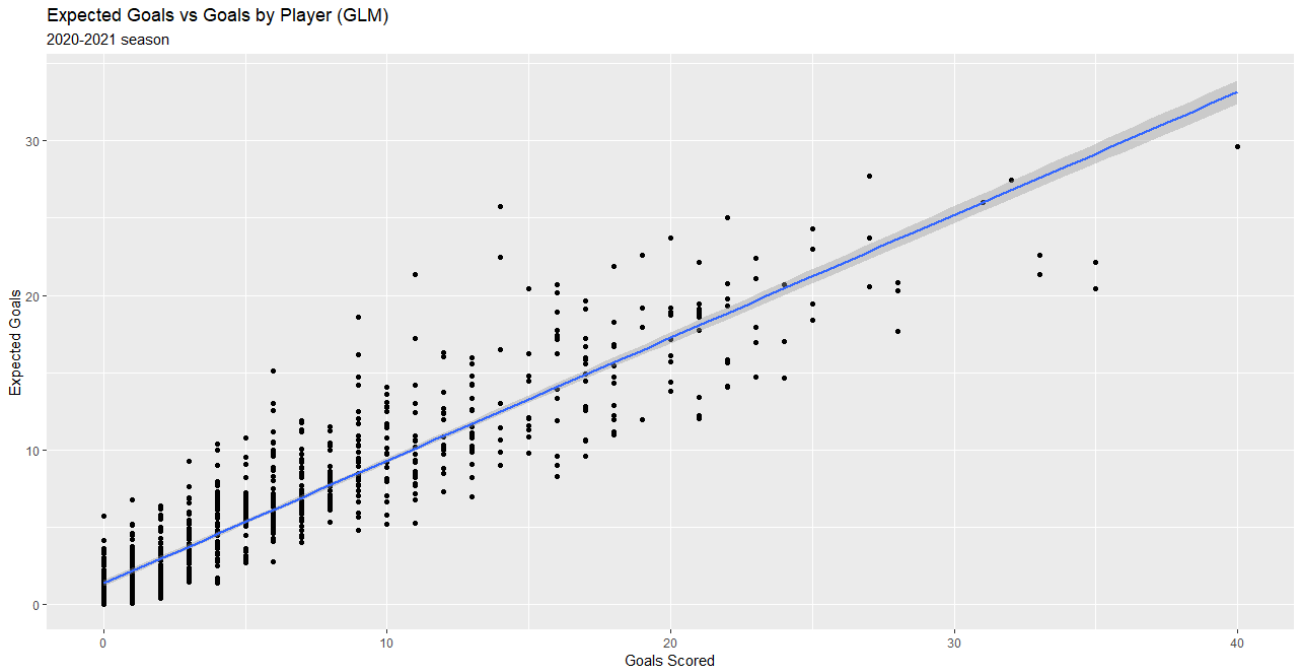

Figure 2: The predicted expected goal total compared to the true goal total for 896 players during the 2020-2021 NHL season

The model is able to deliver a spatial distribution of expected goals that is based on the location of the shot taken as well as the distant from the net independently. Of the locations in the offensive zone, expected goals are predicted to be highest in close proximity to the net with a shot coming from the middle of the rink's width. The expected goals decline as a shot is taken further away from the net and a sharper angle. There are cases where the expected goals are high for a shot that is behind the net or at a great distance from the net. A low number of opportunities come from these locations, deflection occurring in front of the net, or a shot on an empty net are all possible explanations for this phenomenon.

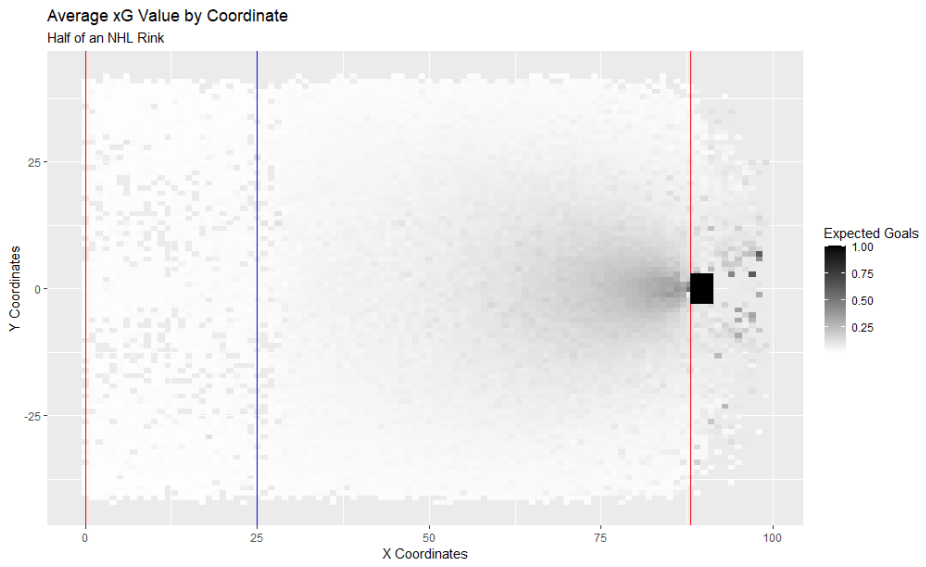

Figure 3: Spatial distribution of predicated expected goals.

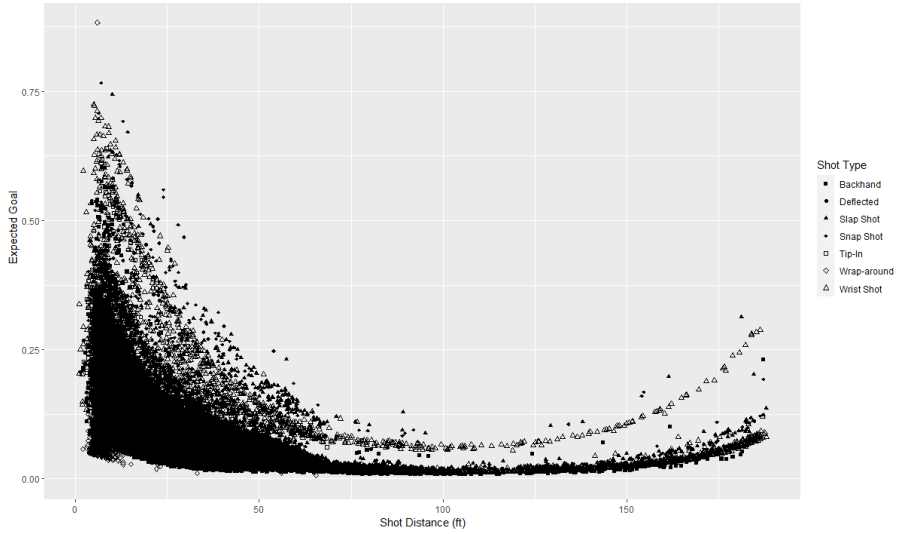

Figure 4: Predicted expected goals based on the distance of shot from the goal.

Predicting NHL games:

Applying the model to an entire NHL game, two methods emerge for examining a team's performance, cumulative expected goals through time and sum of individual contributions. Both of these methods arrive at the same total expected goals per team using difference avenues to reach these results. Cumulative expected goals calculated the sum of expected goals for each team as time passes during the game. This method showcases a temporal outlook on the game and reveals which team is more likely to win based on their position relative to each other. Additionally, this method is capable of illustrating the dominant team in a specific section of the game by comparing the slope of expected goals during that time. A larger slope indicating the stronger team at that moment.

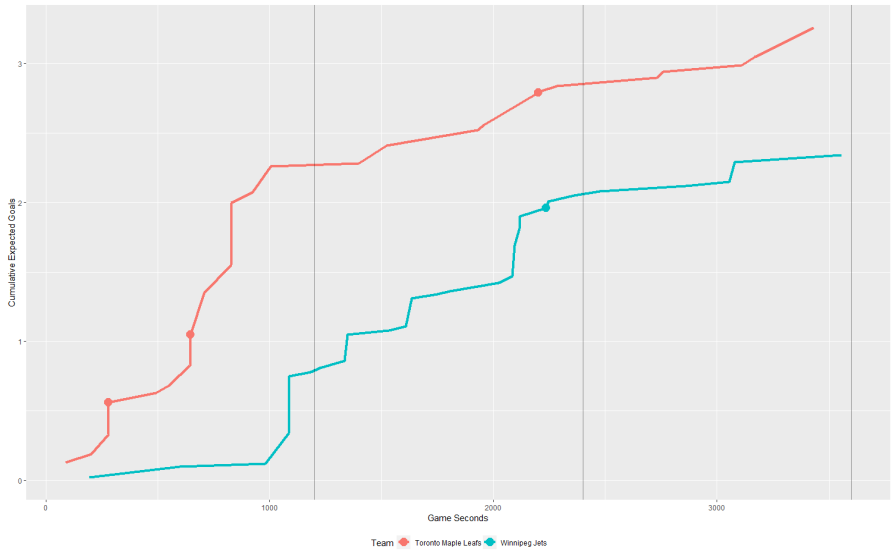

Figure 5: Cumulative expected goals through time for a game of March31st 2021 between the Toronto Maple Leafs and the Winnipeg Jets. Time is indicated in seconds.

The sum of individual contributions calculates the sum of expected goals from each player and then takes the sum of all the players to reach a total expected goals for each team while showcasing the individual contribution of each player. This method delivers a single number for each team to determine the number of expected goals were earned throughout the game, without any temporal indicators. This method also allows us to evaluate each player's contribution to offense in a game and determine who is most valuable to their team. A player who does not appear in the figure did not contribute a shot on net during the game and therefore does not any expected goals

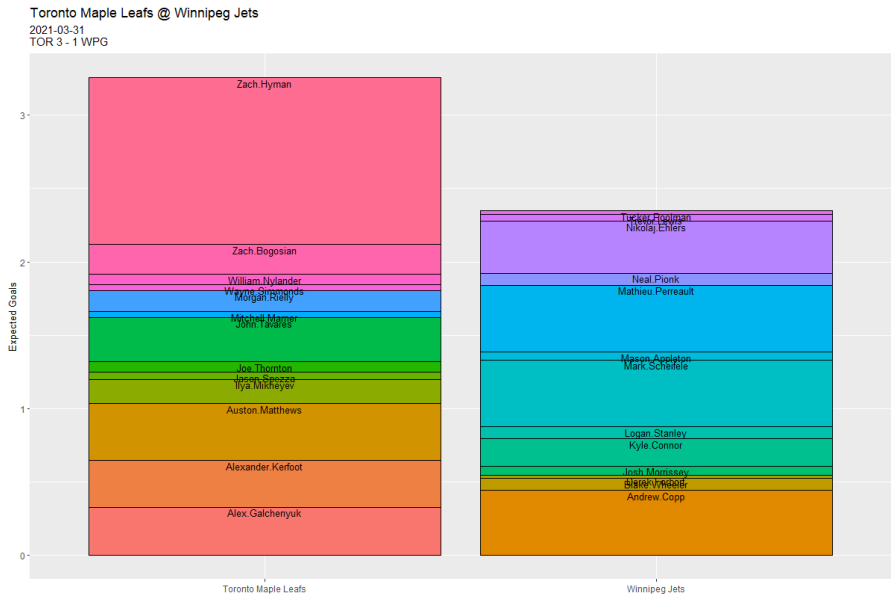

Figure 6: Expected goals per teams based on the sum of individual expected goals for a game of March31st 2021 between the Toronto Maple Leafs and the Winnipeg Jets.

Predicating a season:

The total expected goals calculated from the sum of individual contributions method was used to predict the outcome of all 868 games of the 2020-2021 season. For each game of the season, the total expected goals were calculated and then compared with the true score of the model. The prediction was deemed correct if the team with higher expected goals was the team was ultimately won the game. A second correct scenario was if the difference in expected goals between the two teams was less than 1.3 and the game was tied after overtime. A difference of 1.3 goals was chosen as a conservative reference to the 1.5 goals used in sports betting to indicate a close game.

For the 868 games of the 2020-2021 season, the model correctly predicted the outcome of 548 games, a 63% accuracy rate. Of the 56 games that each of the 31 teams played, the model successful predicted an average of 36 games. The model was the most successful with the Montreal Canadiens and the New York Islanders, correctly predicting the outcome for 42 of the 56 games. The model was the least successful with the New Jersey Devils, correctly predicting only 29 of the 56 games, a 52% accuracy rate. This method was also applied to the training seasons of 2018-2019 & 2019-2020. Of the 2365 games included in the dataset, the model correctly predicted the outcome of 1436 games, a 61% accuracy rate.

These rates of accuracy of similar to existing publicly available expected goals models. Morrison & Rad (2018) used a more complex neural network with 11 hidden layers to successfully predict the outcome of 3720 games between 2015 – 2018, successfully predicting 60% of the outcomes. A paper by Weissbock et al. (2014) argued that a ceiling of 62% accuracy exists due the randomness in hockey that no model would be able to exceed this ceiling. The expected goals model used to predict the outcome of the 868 games of the 2020-2021 season in this analysis was able to exceed the proposed maximum accuracy by 1% with caveat that the criteria for correct predictions did not include a decision when games were decided by a shootout. Overall, the expected goals model was able to predict the outcome of NHL games with over 60% accuracy, yielding similar results to other models with surrounding a theoretical maximum for predictive power.

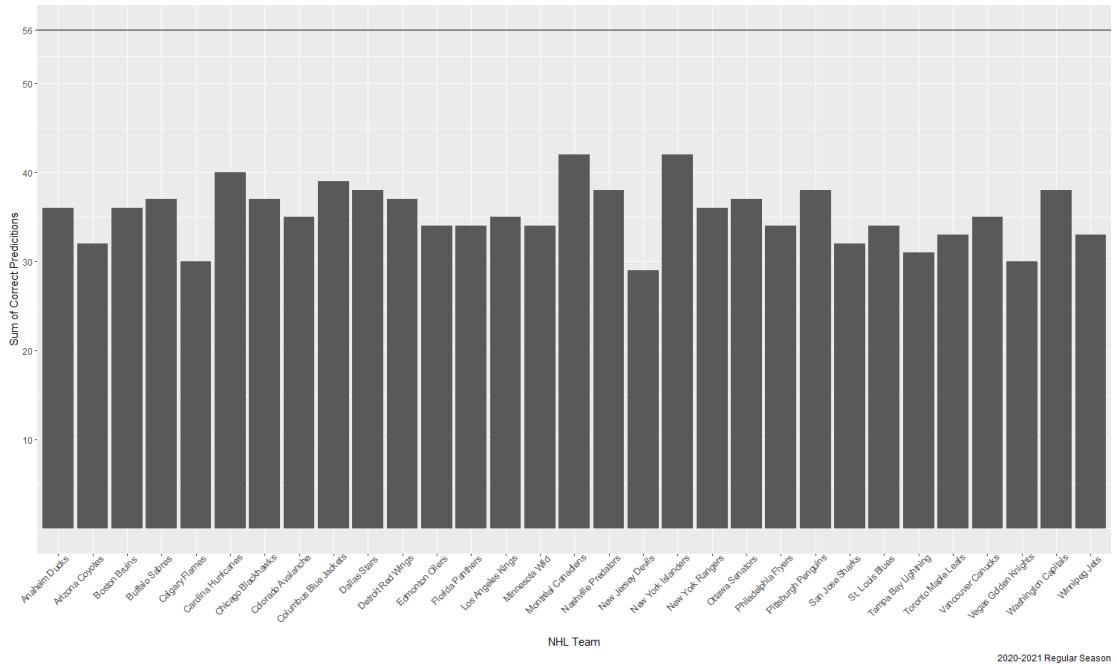

Figure 7: Number of correctly predicted games per team out of 56 total games for the 2020-2021 NHL season.

Figure

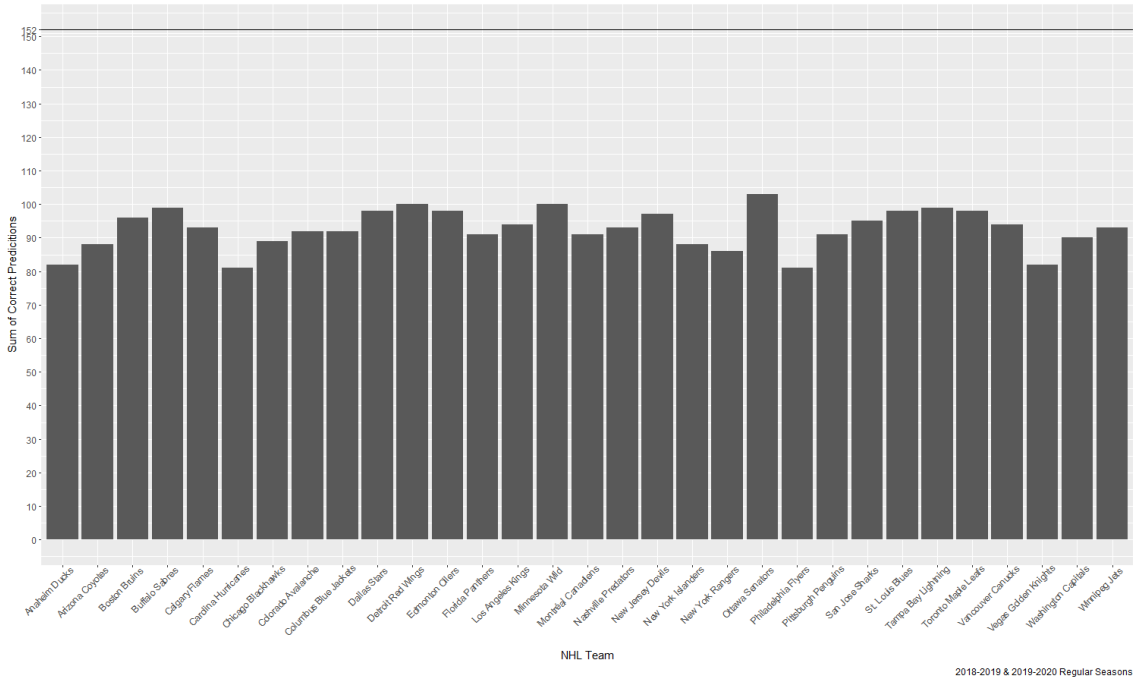

Figure 8: Number of correctly predicted games per team out of 152 total games for the 2018-2019 & 2019 - 2020 NHL seasons.

Players monetary value

For the 2020-2021 season, the maximum budget that teams could spend on players was $81.5 million. A typical hockey roster consists of 23 players. Given these two metrics, a team should spend approximately $3.5 million per player for the season. When examining the salary structures of teams across the league, there are numerous player that exceed this number with player's salaries exceeding $10 million per season. The salary structure of most teams in the NHL pays a hand full of star players larger salaries while filling their remaining roster slots with inexpensive player than are at or near the minimum NHL salary of $700,000 per season. While player salaries vary across the league, so do the contributions of each player. One method to assess a player's financial value is through an analysis of dollars per expected goal. Using the total expected goals for the season and dividing by the cap hit, a unit cost per expected goal can be calculated to understand if a player's contributions are worth the salary they are being paid.

$$Dollars\ per\ xG = \frac{Season\ Total\ xG}{Cap\ Hit}$$

When reviewing the results of the 2020-2021 season, the 16 teams that made the playoffs averaged 180 goals. Using the formal for 180 goals and an $81.5 million payroll, gives a theoretical maximum a team should spend per expected goal of $452,777. When applying this equation to every player, there are is a trend that occurs with three types of players, players who have a high unit cost with low offensive contribution, players who have a low unit cost and a low offensive contribution, and players with a low unit cost with high offensive contribution.

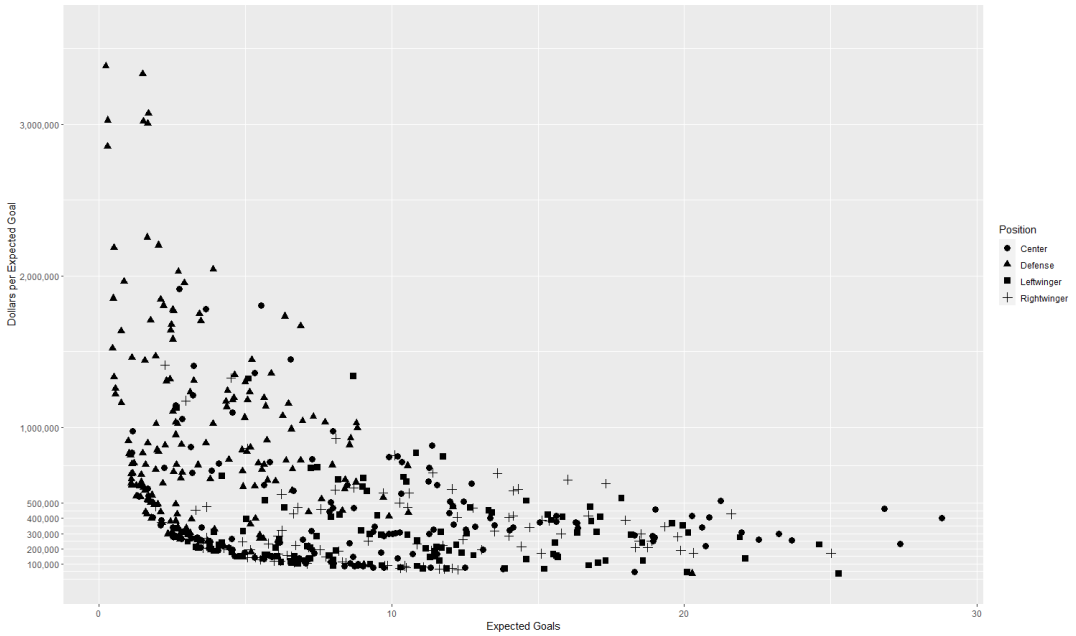

Figure 10: Dollars per expected goals compared with expected goals for the 2020-2021 NHL season.

When examining the players with the lowest dollars per expected goal, there are two types of players that emerge, young players on entry level contracts and aging players on low-cost contracts. The player with the highest expected goals for the season is Auston Matthews with 28.8 expected goals. With a salary of $11.634 million, it is one of the highest in the NHL. His dollars per expected goals is $404,355 per goal, slightly lower than the theoretical maximum. Among all players, Brady Tkachuk of the Ottawa Senators contributed 25.3 expected goals for the season with a salary of $725,000, yielding a dollars per expected goal of $36,604. If a team had that value for all 180 goals needed to reach the playoffs, they would spend $6,588,720 in salary which is only 8.1% of maximum salary budget of $81.5 million.

Table 2: Lowest dollars per expected goal for all types of players for the 2020-2021 season

Player	Age	Team	Expected Goals (xG)	Cap Hit	Dollars per xG	Total Cost for 180 Goals
Brady Tkachuk	20	OTT Senators	25.3	$725,000	$36,604	$6,588,720 (8.1%)
Kirill Kaprizov	23	MIN Wild	20.1	$925,000	$45,995	$8,279,100 (10.1%)
Andrei Svechnikov	20	CAR Hurricanes	20.1	$832,500	$46,001	$8,280,180 (10.1%)
Nick Suzuki	20	MTL Canadiens	18.3	$863,333	$47,213	$8,498,340 (10.4%)
Corey Perry	35	MTL Canadiens	12.3	$750,000	$61,147	$11,006,460 (13.5%)
Conor Garland	24	AZ Coyotes	11.8	$775,000	$64,722	$11,649,960 (14.3%)
Jordan Kyrou	22	STL Blues	11.6	$758,333	$65,179	$11,732,220 (14.3%)
Carter Verhaeghe	24	Florida Panthers	15.2	$1,000,000	$65,696	$11,825,280 (14.5%)
Conor Sheary	28	BUF Sabres	11.1	$735,000	$66,413	$11,954,340 (14.7%)
Joel Farabee	20	PHI Flyers	13.9	$925,000	$66,660	11,998,800 (14.7%)

Because players on entry level contract have a maximum salary and extensive rights protection for the team, a team looking to acquire value players would have to draft good players and allow time for development. These teams could be looking for players that can be acquired to improve their team rapidly. Among the players who have either been traded or signed with another team during their career, Corey Perry of the Montreal Canadiens was the best value player with a dollars per expected goals of $61,147. If a team had that value for all 180 goals needed to reach the playoffs, they would spend $ 11,006,460 in salary which is only 13.5% of maximum salary budget of $81.5 million. Of the ten players listed in table X, Barclay Goodrow and Blake Coleman were both members of the Stanley Cup winning Tampa Bay Lightning and Corey Perry was a member of the Stanley Cup finalists, the Montreal Canadiens.

There is a potential relationship that exists between teams that are able to acquire value players and success during the season.

Table 3: Lowest dollars per expected goal for acquired players for the 2020-2021 season

Player	Age	Team	Expected Goals (xG)	Cap Hit	Dollars per xG	Total Cost for 180 Goals
Frans Neilsen	36	DET Red Wings	2.7	$5,250,000	$1,913,756	$344,476,080 (423%)
Artemi Panarin	28	NY Rangers	8.7	$11,642,857	$1,338,376	$240,907,680 (296%)
Patrik Laine	22	CB Blue Jackets	5.1	$6,750,000	$1,321,512	$237,872,160 (292%)
Valterri Filppula	36	DET Red Wings	2.6	$3,000,000	$1,147,038	$206,466,840 (253%)
Evgeny Kuzetsov	28	WSH Capitals	7.9	$7,800,000	$975,072	$175,512,960 (215%)
Jonathan Drouin	25	MTL Canadiens	6.2	$5,500,000	$893,035	$160,746,300 (197%)
Riley Nash	31	CB Blue Jackets	3.1	$2,750,000	$873,081	$157,154,580 (193%)
Brandon Sutter	31	VAN Canucks	5.1	$4,375,000	$862,091	$155,176,380 (190%)
Jeff Skinner	28	BUF sabres	10.8	$9,000,000	$830,237	$149,442,600 (183%)
Ryan Johanson	27	NSH Predators	9.9	$8,000,000	$806,428	145,157,040 (178%)

Of the players with the highest dollars per expected goal only forwards were assessed due to the potential value of defensively oriented defense that are valuable to their team even if they do not contribute offense. Among these players there are 4 forwards with dollars per expected goals over 1$ million. Frans Neilsen of the Detroit Red Wings was the highest at $1.9 million per expected goal, making him the worst value forward in the league for that season. If a team had that value for all 180 goals needed to reach the playoffs, they would spend $344,476,080 in salary which is 423% or over 4 times the maximum salary budget of $81.5 million.

Table 4: Highest dollars per expected goal for all types of players for the 2020-2021 season

Player	Age	Team	Expected Goals (xG)	Cap Hit	Dollars per xG	Total Cost for 180 Goals
Corey Perry	35	MTL Canadiens	12.3	$750,000	$61,147	$11,006,460 (13.5%)
Carter Verhaeghe	24	Florida Panhers	15.2	$1,000,000	$65,696	$11,825,280 (14.5%)
Conor Sheary	28	BUF Sabres	11.1	$735,000	$66,413	$11,954,340 (14.7%)
Trevor Moore	25	LA Kings	10.5	$775,000	$73,837	$13,290,600 (16.3%)
Jason Spezza	37	TOR Maples Leafs	9.4	$700,000	$74,723	$13,450,140 (16.5%)
Barclay Goodrow	27	TB Lighting	10.9	$925,000	$84.975	$15,295,500 (18.8%)
Nic Dowd	30	WSH Capitals	8.7	$750,000	$85,853	$15,453,540 (18.9%)
Nick Ritchie	24	BOS Bruins	16.7	$1,498,925	$89,552	$16,119,360 (19.8%)
Curtis Lazar	25	BUF/BOS	8.9	$800,000	$90.341	$16,261,380 (19.9%)
Blake Coleman	28	TB Lightning	17.1	$1,800,000	$105,477	$18,985,860 (23.3%)

Teams' money efficiency:

Considering the theoretical maximum of dollars per expected goal of $452,777. Assessing how teams spend their money on players could offer insight into the state of management of each team. For each team, the dollars per expected goals was calculated for each player and the mean was taken to understanding how a team spends their money on average. Figure X also indicates the ranking of team at the end of the season with 1 being the 1st place team and 31 being the last place team. When reviewing the data, only 7 of the 31 teams are at or below the theoretical maximum amount spent per expected goal. Of the teams that are below the theoretical maximum are the teams the finished 1st, 2nd, 4th, 6th, 8th, 10th, and 17th in the regular season standings. Indicates that some of the top performing teams in the league spent the least amount of the per expected goal. On the other side of the standings, the 29th, 30th, and 31st team in the regular season standings ranked among the four teams that spent the most per expected goal. Of the 16 teams that spent the lowest amount per expected goal, 12 of them made the playoffs with 2 narrowly missing. As exemplified in the figure, there is a clear relationship regular season performance and spend a lower amount per expected goal. Additionally, a team is more likely to make the playoffs if they rank in the top half in average dollars per expected goals.

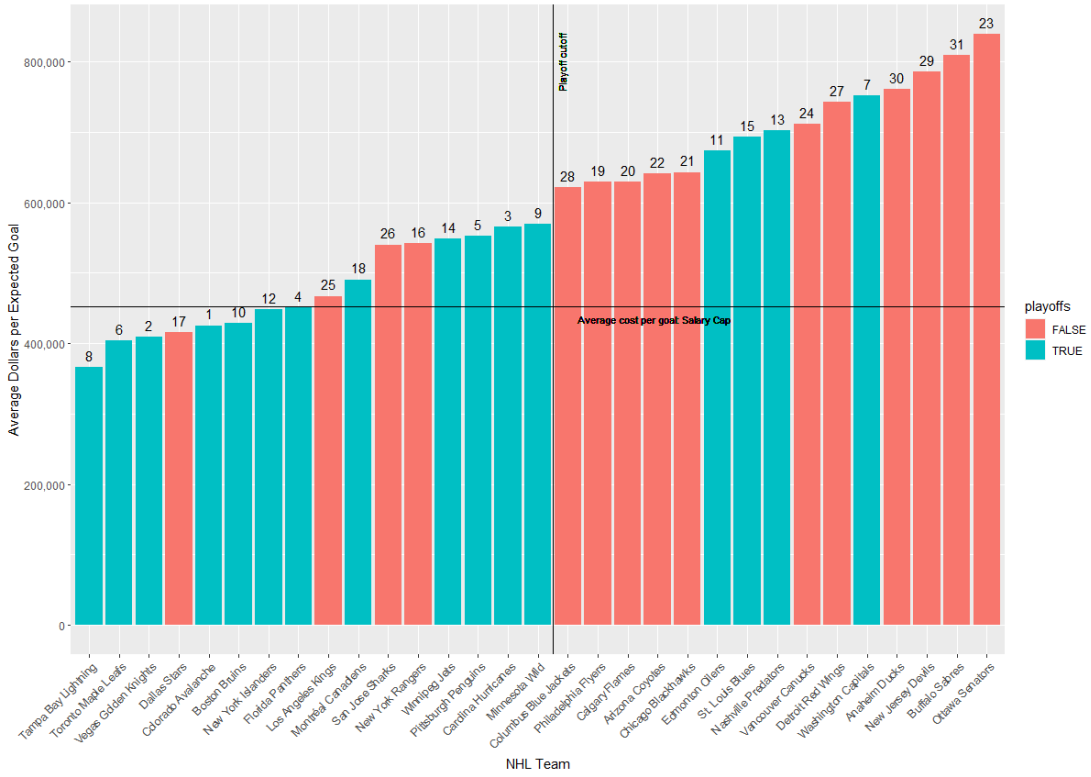

Figure 11: Average dollars per expected goals spent by each team. Final standings for the regular season indicated above the bar. 16 teams with the lowest dollars per expected goals indicated by vertical line. Theoretical maximum of $452,777 indicated by horizontal line.

Conclusion:

Using the open-source play-by-play data provided by the NHL. An expected goals model was developed using logistic regression that consisted of a myriad of positional, situational, and personnel data. The model proved to be highly effective at creating a sum of expected goal that was close to the true goal totals of NHL players. The model was used to evaluate the individual contributions of players during a game and predict the outcome of the game based on the sum of the contributions. This prediction method was used to predict the outcome of all the games of 2018-2019, 2019-2020, and 2020-2021 NHL seasons with a prediction accuracy ranging between 61-63%. The predicted expected goals were paired with NHL salary data to analyze the monetary value of players based on their contribution to offense yielding a metric called dollars per expected goals. The values for each player were aggregated to the team level to evaluate the financial efficacy of each team in relation to their in-season performance. Overall, the logistic regression model using open-source data was highly effective at describing NHL games and individual contributions with the ability to add additional variables for deeper analysis

References:

Liu, G., Schulte, O., Poupart, P., Rudd, M., Javan, M. (2020) Learning Agent Representations for Ice Hockey. *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

MacDonald, B. (2012) An Expected Goals Model for Evaluating NHL Teams and Players, *MIT Sloan Sports Analytics Conference.*

Douglas, E., Clement, S., Wan, N., Greengross. (2021) Valuing Individual Contributing Events (V-ICE) in Hockey.

Morrison, J. & Rad, N.F. (2018) A Machine Learning Approach to Predicting Regular Season Success in the National Hockey League. *CPSC.*

Gramacy, R.B., Hensen, S.T., Taddy, M., (2013) Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports.* 9(1).

Weissbock , J. (2014) Forecasting Success in the National Hockey League Using In-game Statistics and Textual Data. *uO Research.*

Marek, P., Sediva, B., Toupal, T. (2014) Modeling and prediction of ice hockey match results. *Journal of Quantitative Analysis in Sports.* 10(3).

Pichedda, G. (2014) Predicting NHL Match Outcome with ML Models. (2014) *International Journal of Computer Applications. 101(9), 15-22.*

Nandakumar, N. & Jensen, S.T. *(*2019) Historical Perspective and Current Direction in Hockey Analytics. *Annual Review of Statistics and its Application,* 6, 19-36.

Luszczyszyn, L. (2016) Measuring Single Game Productivity: An Introduction to Game Score. *Hockey-graphs.com*

Idson, T.L. & Kahane. L.H (2007). Team effects on conpensation: an application to salary determination in the National Hockey League. *Economic Inquiry,* 38(2), 345-357.

Weissbock, J., & Inkpen, D. (2014). Combining Textual Pre-game Reports and Statistical Data for Predicting Success in the National Hockey League. Canadian Conference on Artificial Intelligence 2014, 8436, 251-262.

Morse, D. (2022). hockeyR: Collect and Clean Hockey Stats. R package version 0.1.0.9000.

National Hockey League. (2013) Collective Bargaining Agreement.

Cox, D.R. (1958) The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*

Peng, C.J., Lee, K.L., Ingersoll, G.M. (2002) An introduction to Logistic Regression Analysis and Reporting. *Journal of Educational Research,* 96(1) 3-14.

Kuss, O., (2002) Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine,* 21(24), 3789-3801.

Bradley, A.P (1997) The use of area under the ROC curve in the evaluation of machine learning algorithms.