NHL Points Predictor

Martin Knuth
University of Pittsburgh
Python for Data Management and Analytics

Table of Contents

I. Abstract

Almost all NHL teams have dedicated departments for analytics research and statistical analysis with the goal of creating models to predict future success. Millions of dollars can hinge on the wrong move or the wrong thought process for teams. Getting the slightest edge on your opponent can be the difference between a championship season and missing the playoffs.

The goal of my model is to predict team success based on statistics that measure many different aspects of NHL teams. Those aspects include goaltending, luck, shooting talent, possession of the puck, shot quality and special teams play. I chose not to use statistics like wins, losses and goal differential because these potential features strongly correlate with points and would ruin the predictive aspect of the model.

My plan for gathering data was to combine datasets from the last ten seasons into one large dataset. The data was gathered from two websites: *Hockey-Reference* and *Evolving-Hockey*. *Hockey-Reference* contained the basic statistics while *Evolving-Hockey* contained shot and shot location data adjusted for the score and venue of the game. After cleaning the data, I used a Random Forest Regressor to estimate team points.

After running the model on the entire set of team data, I found that my R^2 value and RMSE value to look very promising. Given these metrics of accuracy, I was very happy with the predictive nature of my model, given how it was set up. On the other hand, I began to ponder the issues with the model and discussed why this model is not as great as the metrics would suggest.

II. Introduction

In the NHL, teams gain 2 points for a win, 1 point for a loss in overtime and none for a loss in regular play with the average cutoff for making the playoffs being 92 points. The specific importance of my model's predictions would be for the purpose of in-season moves. Around late February a deadline for trades league-wide is enacted. This deadline creates two types of teams: buyers and sellers. Buyers are teams, typically in playoff spots at the time, looking to add players via trade before the deadline. Selling teams are the teams with no hope of making the playoffs who will trade-off pending free agents for younger players or draft picks. This is where my model would come into play: teams, part-way through the season, would be able to predict where they would finish in the standings to see if they should be selling players off for future assets or loading up for a playoff push.

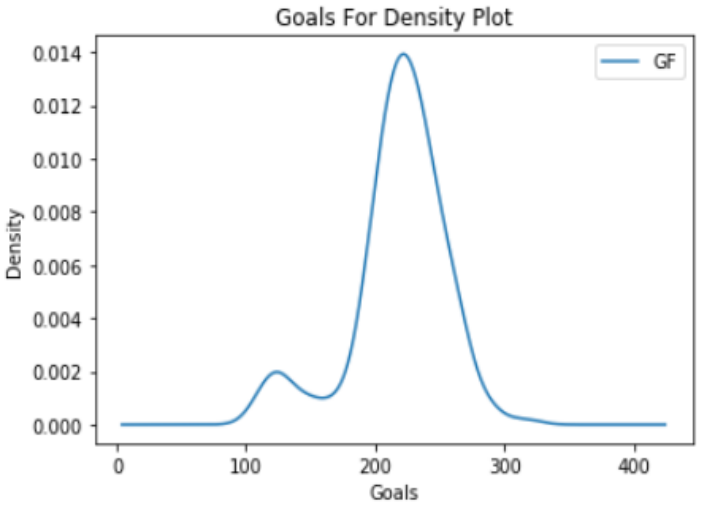
III. Methodology

My original plan was to use a classification model to classify teams as a playoff team or not, but after some digging into the data, I switched to a regression model to predict team points. The reason for this was because of the NHL's playoff format. The NHL is currently split into two conferences, the East and West, each with two divisions. This is shown below:

Western Conference		***Eastern Conference***	
Pacific Division	**Central Division**	**Metropolitan Division**	**Atlantic Division**
Anaheim Ducks	Chicago Blackhawks	Carolina Hurricanes	Boston Bruins
Arizona Coyotes	Colorado Avalanche	Columbus Blue Jackets	Buffalo Sabres
Calgary Flames	Dallas Stars	New Jersey Devils	Detroit Red Wings
Edmonton Oilers	Minnesota Wild	New York Islanders	Florida Panthers
Los Angeles Kings	Nashville Predators	New York Rangers	Montreal Canadiens
San Jose Sharks	St. Louis Blues	Philadelphia Flyers	Ottawa Senators
Vancouver Canucks	Winnipeg Jets	Pittsburgh Penguins	Tampa Bay Lightning
Vegas Golden Knights		Washington Capitals	Toronto Maple Leafs

This presented some challenges. The biggest was that, although 16 teams make the playoffs every year, each conference is allotted 8 playoff spots. This presented issues because 10 teams in my dataset have finished in the top 16 but did not make the playoffs. For example, the 2018-19 Montreal Canadiens, who did not make the playoffs, finished in 14th place in total points which put them ahead of 3 teams who did make the playoffs. In that year, the Eastern Conference had better teams, so the Canadiens missed the playoffs because all of the spots allocated for the East were filled. As a result, one could argue that a variable for what conference a team is in should be put into the dataset. This would not help because the playoff format has fairly recently been changed and the strength of a conference in each year differs. The best solution I've found was to change my response variable to the common denominator in each season: team points. If I can predict how many points all the teams in a season have, I can use those predictions to fill in the playoff format for that given year, thus avoiding this issue.

The next issue I ran into was the 2012-13 shortened season. Due to a player-owner dispute that bled into the season, a full season was not played. This presents issues to stats that use totals rather than rates. This can be illustrated in the density plot of goals scored for each team. The bimodal natural of this plot is a symptom of this shortened season. Had teams been able to play the regular 82 game season, this plot would be a normal distribution. The solution for this issue was to rate each stat based on games played or minutes in that situation.



For the shortened season, all the teams point totals were lower than they should be because they didn't play the full 82 games. To fix this I created a new data point, Points Per 82 Games Played. PTS/82 extrapolates the pace each team was on to a full season. The formula for PTS/82 is:

(total points / (games played * 2)) * (82 * 2) = PTS/82
(percentage of points earned) * (points possible in an 82-game season) = PTS/82

After some cleaning and combining of two datasets into one, I finally began feature selection. Between both datasets, I had 54 features to choose from. I was able to remove the vast majority of them without much thought because they were either irrelevant or too similar to the response variable (like wins and losses). After some deliberation, I brought the list of features down to a reasonable amount of 8:

Feature	Reason for consideration
Powerplay Percentage (PP%)	A measure of powerplay successes to total powerplays. Seasons can be won or lost due to a very effective or ineffective powerplay. (A powerplay is when at least one opposing player is serving a penalty, and the team has a numerical advantage on the ice)
Penalty Kill Percentage (PK%)	A measure of penalty kill successes to total penalty kills. (A penalty kill is an event when a team that is minus one or more players due to a penalty, also referred to as "short-handed")
Team Save Percentage at All Situations (SV%)	Percentage of shots saved by the team's goalie at all situations, whether at even strength, the powerplay or penalty kill.
Team Shooting Percentage at All Situations (S%)	Percentage of shots converted into goals at all situations.
Corsi For Per 60 Minutes (CF/60)	Corsi is a measure of shot attempts while at even strength play. This includes shots on goal, missed shots on goal, and blocked shot attempts towards the opposition's net.
Corsi Against Per 60 Minutes (CA/60)	Corsi against at even strength per 60 minutes of ice time.
Expected Goals For Per 60 Minutes (xGF/60)	Expected goals is a measure of shot quality. Given certain characteristics about a shot, xGF will predict the likelihood of that shot becoming a goal. Those characteristics include: • Shot type (Wrist shot, slap shot, deflection, etc.) • Shot distance (Adjusted distance from the net)

	• Shot angle (Angle in absolute degrees from the central line normal to the goal line) • Rebounds (Whether or not the shot was a rebound) • Rush shots (Whether or not the shot was a rush shot)
Expected Goals Against Per 60 Minutes (xGA/60)	Expected goals against per 60 minutes of even strength time.

Some notable features I considered but did use were:

- Shots For and Against: covered under CF/60 and CA/60 in a more efficient way
- Goals For and Against: goal differential is too similar to wins
- Fenwick For and Against: Fenwick is a measure of unblocked shot attempts. So while Corsi will include shot attempts that are blocked, Fenwick will not. I made the decision to use Corsi over Fenwick as a measure of processing the puck in the offensive zone.
- Even Strength Shooting and Save Percentage: covered under all situation SV% and S%

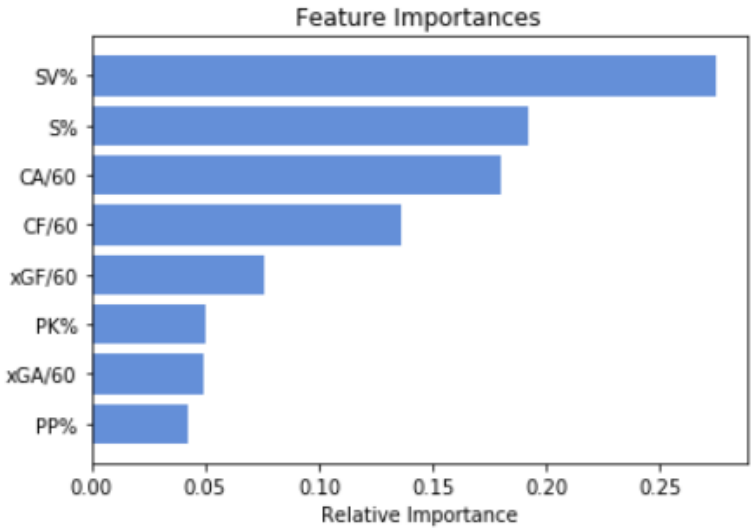
One more important note for feature selection: I dropped the rates of Corsi and Expected Goals (also known as CF% and xGF%) to use CF/60, CA/60, xGF/60 and xGA/60. This was because I wanted to see if creating shot attempts or preventing shot attempts was more important. For example: Team A could produce 100 shot attempts per game on average while allowing 100 shot attempts against. Team A's CF% would be 50%, because it allows the same amount of shot attempts as it gives up. Team B could produce 10 shot attempts per game on average while allowing 10 shot attempts on average against. Both Team A and B would have CF% of 50. By using a percentage of attempts for, I would have been unable to determine if producing or allowing shot attempts is more important.

After cleaning the data, I decided to use a Random Forest Regressor for my machine learning algorithm. I chose this algorithm because of its use of supervised machine learning and the size of the dataset did not present high computational costs. Random Forest would also allow me to analyze the feature importance. I tested some other algorithms like SVM and Linear Regression but my original assumption of using Random Forest turned out to be the best algorithm.

III. Results

I was pleasantly surprised with the results of my model. Given full seasons of team data, my model was able to predict the point total of a team within an average of 1.86 points, which is less than one win in regulation. The R^2 value of my model was .9707 which means that 97.07% of the variability in team points can be explained using my model. The root mean squared error value (RMSE) is 2.488. This low RMSE bodes well for the predictive value of my model.

The feature importance chart, on the right, tells an interesting story. Save percentage being the most important feature does not surprise me, because bad goaltending can tank a good team and good goaltending can carry a bad one. To prove this, only 32.35% of teams that missed the playoffs had an above average save percentage, while 70.83% of playoff teams had an above average save percentage. All other features being average, a team with an average save percentage will have 8.01 predicted points more than a team with a save percentage one standard deviation less than the average.



To further illustrate the importance of goaltending, I made a new dataset of all the teams in my original dataset but gave every team the league average save percentage of 90.8%. I reran the prediction model on this new dataset to see what teams gained or lost the most amount of points due to goaltending in a single season. Here is a list of the five teams that gained the most points in the standings due to goaltending:

Team	Season	Predicted_Points	Avg SV Predicted Pts	PTS/82	SV%	Difference in predicted points
Montreal Canadiens*	14-15	105.394000	93.029500	110.000000	0.926	12.364500
Chicago Blackhawks*	12-13	124.055250	112.101542	131.541667	0.923	11.953708
Colorado Avalanche*	13-14	107.392000	96.423542	112.000000	0.919	10.968458
New York Rangers*	14-15	110.049458	101.107875	113.000000	0.923	8.941583
Washington Capitals*	16-17	116.686000	108.072458	118.000000	0.922	8.613542

The team at the top of the list, the 2014-15 Montreal Canadiens, gained 12.365 predicted points because of Carey Price's incredible MVP winning season where he posted a 93.3% save percentage. The next team on that list, the 2012-13 Chicago Blackhawks, didn't get a chance to fully realize their earned 12 points in the standings due to goaltending because this was the shortened season, but they did win a Stanley Cup that year.

Here is the list of the five teams that lost the most amount of points due to poor goaltending. A couple of these teams, namely the 2012-13 Tampa Bay Lightning and the 2018-19 Ottawa Senators, would have been decent had their goalies not been as bad as they were.

Team	Season	Predicted_Points	Avg SV Predicted Pts	PTS/82	SV%	Difference in predicted points
Tampa Bay Lightning	12-13	73.193125	94.393125	68.333333	0.899	-21.200000
Ottawa Senators	18-19	68.552292	88.965000	64.000000	0.897	-20.412708
Calgary Flames	12-13	74.263583	92.716250	71.750000	0.889	-18.452667
Colorado Avalanche	16-17	55.699583	70.786292	48.000000	0.894	-15.086708
Colorado Avalanche	11-12	71.017625	85.986625	68.000000	0.890	-14.969000

While I still have the average goaltending dataset, let's look at what the best rosters over the last 11 seasons have been and their post season results. This is a list of the top ten teams ordered by predicted points with all teams' save percentage being average (in other words, the teams with the best rosters not including the goalies).

Team	Season	PTS/82	Avg SV Predicted Pts	Post Season Results
Tampa Bay Lightning*	18-19	128.000000	117.113500	Swept in the first round
Washington Capitals*	09-10	121.000000	115.076125	Lost in 1st round
Chicago Blackhawks*	12-13	131.541667	112.101542	Won Stanley Cup
Pittsburgh Penguins*	12-13	123.000000	110.832958	Lost in Conference Finals
Winnipeg Jets*	17-18	114.000000	109.395250	Lost in Conference Finals
Boston Bruins*	17-18	112.000000	109.286708	Lost in 2nd Round
Boston Bruins*	13-14	117.000000	108.917083	Lost in 2nd Round
Detroit Red Wings*	08-09	115.000000	108.911208	Won Stanley Cup
Washington Capitals*	15-16	120.000000	108.870458	Lost in 2nd Round
Pittsburgh Penguins*	11-12	108.000000	108.719458	Lost in 1st round

The next feature in terms of importance for my model was shooting percentage, the number of shots that are converted to goals. There's been debates in the NHL analytics community whether shooting percentage is a measure of shooting talent or luck. My opinion, as well as the opinion of most analysts now, is that shooting percentage is a combination of luck and shooting talent. The best way to establish this combination of skill and luck is, again, to look at the top five teams sorted by their shooting percentage. The first two teams, the 2018-19 Lightning and the 2009-10 Capitals, are two of the most skilled teams in terms of shooting talent from the last 11 years. They boast players throughout their roster with incredible shots, so it's no surprise to see these two teams at the top of the list. The last three teams on this list all have one thing in common: they played in the 2012-13 shortened season. My belief is that because they

Team	Season	S%
Tampa Bay Lightning*	18-19	12.2
Washington Capitals*	09-10	11.6
Toronto Maple Leafs*	12-13	11.5
Pittsburgh Penguins*	12-13	11.3
Tampa Bay Lightning	12-13	11.1

played a shortened season of 48 games, they were able to maintain a higher than normal shooting percentage because the season wasn't long enough for them to regress closer to the mean. Had the season been longer, their luck would have run out eventually and they wouldn't have been at the top of this list. This goes to show that shooting percentage is a measure of luck and skill, a strong feature in predicting success.

I'll group the shot data together because they tell similar stories. I was a little surprised with the ordering of CA/60 being ahead of CF/60 and xGF/60 being less important than xGA/60. This could be random chance or what I've gathered from the ordering: preventing offensive zone time is slightly more important than creating it for your team and that creating quality chances is more important than preventing them.

The last thing I wanted to touch on with the shot data was the adjustment that I chose to use. *Evolving-Hockey* offers data that is adjusted for the score and venue of the game. What this does is weight each shot attempt based on the state of the game. So, for example, a shot taken by a team down by a goal in a visiting arena is weighted more than a shot taken at their home arena up by 6 goals. I believe this was important to include so each shot is weighted closer to its true value.

The last two features relate to each other because they are the only features that consider special teams play. Being a good even strength team is very important but, like goaltending, special teams play can make or break teams.

Team	Season	Points Gained
San Jose Sharks*	09-10	8.674708
Chicago Blackhawks*	12-13	7.532500
Tampa Bay Lightning*	18-19	7.500958
Washington Capitals*	15-16	5.922750
Vancouver Canucks*	11-12	4.962208

Team	Season	Points Gained
Colorado Avalanche	16-17	-7.978250
Florida Panthers	11-12	-7.602958
Ottawa Senators	15-16	-7.453208
Columbus Blue Jackets	11-12	-7.194167
Florida Panthers	13-14	-6.486708

These are the top 5 teams that gained and lost the most points in a season because of their specials teams. The method obtained to gather this data was the same method used to find points gained due to save percentage. I changed every team's special team's percentages to the mean PP% and PK%, ran the model on this data and looked at the difference in predicted points.

Considering how many features I began with, I believe I found a way to incorporate every aspect of a team's success in my model with the fewest dimensions and that every feature I included added value to the predictive nature.

IV. Discussion

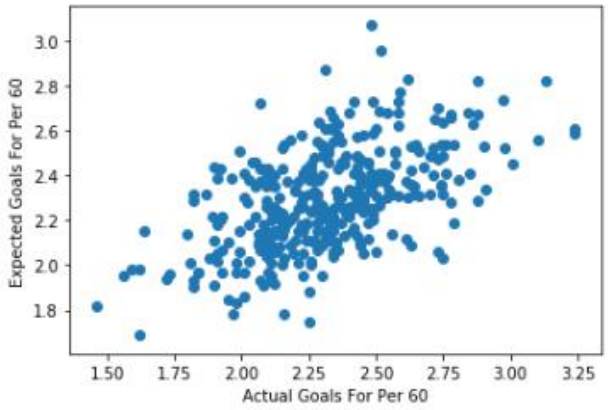
I've discussed the intricacies of the features and results of my model, but I have yet to dissect the issues that hold it back. There are three major issues I see with the model that I was unable to mitigate with my methodology: insufficient sample sizes, the changing of the rosters and issues with expected goals.

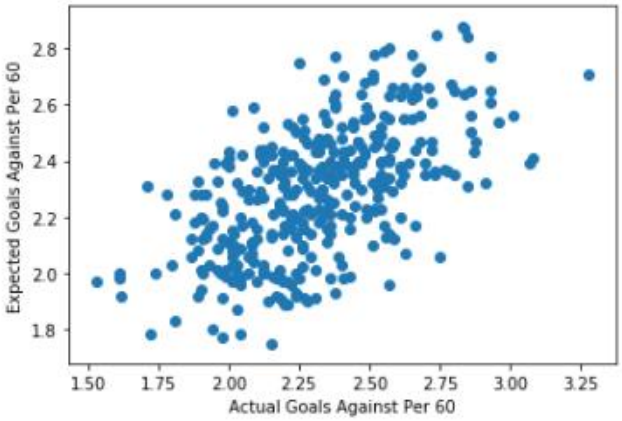
Name	Predicted_Points	PTS/82
Capitals	114.296292	128.466667
Bruins	112.263375	131.785714
Canucks	106.800583	90.482759
Avalanche	106.780500	109.333333
Blues	105.301583	114.800000
Hurricanes	102.882500	96.642857
Flyers	102.189042	108.357143
Penguins	101.901833	99.571429
Islanders	101.309083	119.846154
Lightning	99.530208	95.120000
Golden Knights	98.232208	92.933333
Oilers	96.394167	101.133333
Stars	95.684833	93.310345
Coyotes	94.958125	101.793103
Panthers	94.825375	94.148148
Predators	93.418083	88.074074
Sabres	92.531125	90.785714
Canadiens	92.525208	87.857143
Maple Leafs	88.800000	82.000000
Jets	87.145250	102.500000
Sharks	85.632750	87.655172
Blackhawks	84.948625	75.925926
Wild	84.726292	87.857143
Rangers	84.148708	91.461538
Ducks	84.024750	82.000000
Flames	82.218208	84.827586
Senators	80.056542	70.689655
Blue Jackets	76.567083	78.962963
Kings	73.418542	67.862069
Devils	67.645750	66.814815
Red Wings	60.450333	46.466667

The first issue I see with the model is that it poorly predicts teams with small sample sizes of games. I ran the prediction the on left with each team having played about 30 games as of December 5th, 2019. The issue lies in the fact that teams can get off to hot starts that the model will assume will last throughout the season. A sample size of 30 games maybe enough games to roughly predict where teams will finish because most of these predictions are close to the pace that teams are on, but the model can't be used on smaller sample sizes because it will assume these trends will last for the entire season.

The other issue that I did not account for that other, more sophisticated models, account for is the analysis of the players on the roster. Let's say a team has very good shot and special team's metrics, but their goalie is terrible. My model can account for this by looking at the team's save percentage. The issue is that if the team acquires a better goalie, the model doesn't change to account for this until the team's average save percentage increases. So while the team's save percentage for the season can still be below average, their current situation is much better. The best approach would be to use a combination of team statistics and a measure of each player's individual impact they have on the team similar to baseball's Wins Above Replacement (WAR). This approach could account for roster moves of players.

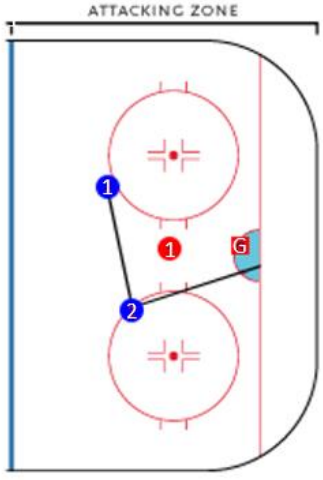
The last issue with the model was that expected goals is still in its infancy as a predictor for goals. In terms of data available to the public, it's the best measure of shot quality, but I wouldn't by any measure call it great:






```
xGF vs GF R^2 value of:  0.2549807571187379
xGA vs GA R^2 value of:  0.29574411359007813
```

The issue, I believe, lies in the lack of player tracking, puck tracking and shooter evaluation. I'll try to illustrate the main issue using the illustration below. This depicts the classic 2 on 1 attack off the rush (2 forwards vs 1 defenseman and the goalie). Attacker 1 passes the puck to attacker 2 for a one time shot into the empty side of the net. The way the current xG model is written this shot would have a relatively low value because of its distance from the net. With player and puck tracking, this shot would have a noticeably higher xG value because, for each shot, the model would be able to measure the following:



- Type of Attack (2 on 1, Breakaway, 3 on 2, 5 on 5, etc.)
- Shooting talent
- Location of the shot on the net (top right corner, bottom left corner, etc.)
- Location of the goalie at the time of the shot
- Speed of the shot
- Distance and angle from the net

So while I was happy with the metrics measuring my model's predictive power, I understand its many limitations. This regression model is by no means perfect, but I believe in the right context it can accurately predict NHL team success.

V. References

"NHL Stats, History, Scores, & Records." Hockey Reference, www.hockey-reference.com/.

Perry, Emmanuel. "Shot Quality and Expected Goals: Part I." *Corsica*, 3 Mar. 2016, www.corsica.hockey/blog/2016/03/03/shot-quality-and-expected-goals-part-i/.

Perry, Emmanuel. "Adjustments Explained." *Corsica*, 19 June 2016, http://www.corsica.hockey/blog/2016/06/19/adjustments-explained/

"Your home for the most advanced NHL statistics on the web!" Evolving-Hockey, www.evolving-hockey.com/.