Kirk Van Arkel
STAT 5010
Project Report
Professor Brian Zaharatos

NHL Predictive Analytics

Introduction

Technological advancements are becoming increasingly prevalent in professional sports each year. From the abundance of high-resolution cameras that allow viewers and referees to track an athlete's every movement, to technologies and devices that autonomously gather statistics of teams and athletes throughout games. The utilization of statistics in professional sports is currently on the rise, bursting on to the scene in recent years following the 2003 book "Moneyball," by Michael Lewis, which documented Billy Beane's successes using sabermetrics as a general manager in Major League Baseball (MLB). Statistics have been a staple in the MLB for more than a decade, but it is only within the past couple years that statistics have become prevalent in other professional sports.

The use of statistics in professional hockey, particularly the National Hockey League (NHL), is not nearly as established as it is in the MLB. There are simply not nearly as many credible advanced statistics in hockey as there are in baseball. The following analyses aim to answer two questions regarding statistics in the NHL:

1) What are the most important hockey statistics in determining an NHL team's regular season success?
2) How accurately can the outcome of an NHL regular season game be predicted strictly based on the two teams' statistics over the previous 15 games?

NHL regular season data from 2008-2019 will be analyzed to answer these two questions. However, data from the lockout season (2012-2013) where only 48 games were played per team (as opposed to the NHL standard of 82 games) will be excluded from these analyses. Many statistics will be considered ranging from easily observable measures such as shots on goal, turnovers, takeaways, and goals, all the way to advanced statistics such as Corsi and Fenwick percentages. There may be some correlation in the variables, particularly shots and Corsi percentage, as Corsi percentage is calculated directly considering shot attempts for and shot attempts against. However, it's difficult to determine other relationships until the data are thoroughly explored.

Exploratory Data Analysis

The data for these analyses is obtained from known NHL analytics expert Peter Tanner at moneypuck.com. The initial acquired spread sheet contained team statistics for every NHL game dating back to the beginning of the 2008-2009 season, where each row stored all the game data of one team for one game. The file contained about 150 different team statistics for each game, many of them being advanced statistics calculated with complicated algorithms. The

dataset was simplified and only 19 statistics were considered for these analyses. A snapshot of the correlations between these 19 statistics can be seen below.

	Avg_SOG_For	Avg_Shot_Att_For	Avg_Goals_For	Avg_Rbds_For	Avg_PIM_For	Avg_FO_Win	Avg_FO_Lost	Avg_Hits_For
Avg_SOG_For	1.00000000	0.88767548	0.577933053	0.45286917	-0.19203501	0.3903669957	-0.070697700	-0.02792015
Avg_Shot_Att_For	0.88767548	1.00000000	0.476536124	0.50599103	-0.20508258	0.4722883291	-0.039637199	0.07386551
Avg_Goals_For	0.57793305	0.47653612	1.000000000	0.16705436	-0.04053884	0.2089074520	0.009519019	-0.14622800
Avg_Rbds_For	0.45286917	0.50599103	0.167054364	1.00000000	-0.20304470	0.3024886447	0.026544396	0.09565062
Avg_PIM_For	-0.19203501	-0.20508258	-0.040538842	-0.20304470	1.00000000	-0.2479624777	-0.208252490	0.12011092
Avg_FO_Win	0.39036700	0.47228833	0.208907452	0.30248864	-0.24796248	1.0000000000	-0.056068142	0.05810322
Avg_FO_Lost	-0.07069770	-0.03963720	0.009519019	0.02654440	-0.20825249	-0.0560681416	1.000000000	0.13605025
Avg_Hits_For	-0.02792015	0.07386551	-0.146227999	0.09565062	0.12011092	0.0581032194	0.136050247	1.00000000

Figure 1: Correlation Matrix showing correlations between 8 of the 19 statistics used for analysis.

To no surprise, there was heavy correlation between shots on goal and shot attempts as well as shots on goal and goals, as seen in Figure 1. There was also heavy correlation between shots on goal, shot attempts and Corsi percentage, however that cannot be seen in Figure 1's snapshot. While the correlation between some statistics was clearly evident, it was worth constructing some models with all of the predictors to help determine which correlated variables had more significant predictive power.

Determining Regular Season Success

To discover the most influential statistics in determining a team's regular season success, each team's average of all 19 statistics over all 82 games in a season were used. These 19 predictors were calculated in attempt to predict the number of regulation/overtime wins (ROW) a team would garner in the 82-game season. Although ROW excludes shootout victories, it is important because it is the first tie-breaker in determining which teams make the playoffs.

After calculating team/season averages, there were 1,205 observations to analyze. These data were randomly split in to training and testing sets, where 80% were used as training data to create a model and the other 20% as a testing set to test the model. Regression techniques utilized included linear regression and binomial regression. With each of the two regression techniques, two different model selection techniques were considered, backward elimination and Akaike Information Criterion (AIC). With each regression technique, both model selection techniques provided the same model, both having 6 predictors, albeit different predictors.

Regression	Model Selection	MSE
Linear	Backward Elimination	5.35
	AIC	5.35
Binomial	Backward Elimination	5.84
	AIC	5.84

Figure 2: Resulting MSEs of the 2 regression techniques with 2 model selection techniques.

Surprisingly, looking at the MSEs in Figure 2, the linear regression proved superior in predicting ROW when compared to the binomial regression. The final model given by the linear regression is:

$$Y = 40.47 + 16.14X_{GF} - 0.36X_{PIM} - 0.29X_{SOG.Ag} - 13.60X_{GA} + 2.19X_{RB.Ag} - 0.15X_{Hits.Ag}$$

Where X_{GF} is the team's average goals scored per game throughout the season, X_{PIM} is the average penalty minutes taken per game, $X_{SOG.Ag}$ is the average shots on goal against, X_{GA} is average goals against, $X_{RB.Ag}$ average rebounds against, and $X_{Hits.Ag}$ average hits against. Looking at coefficient values, it's clear that average goals for and average goals against are extremely influential in predicting number of wins, with values of 16.14 and -13.60, respectively. This means that an increase in 1 average goal for per game would on average lead to 16 more regulation/overtime wins in a season, which is a large impact. However, a one goal increase in average goals per game means the team would have to score 82 more goals throughout the duration of the season, which is not an easy task.

Team	Season	Number.of.Wins	Predicted.Wins
FLA	2017	41	38
DET	2014	39	40
CAR	2010	35	37
FLA	2008	38	38
NYR	2009	35	35
T.B	2013	38	41

Figure 3: Six example predictions of ROW from the model evaluated on the testing data.

Looking at Figure 3, the model looks quite accurate, and an MSE of 5.35 means the model is on average only off by about 2 wins, a respectable result. Using a linear model on this count data, it was important to ensure the model satisfied the linear model assumptions.

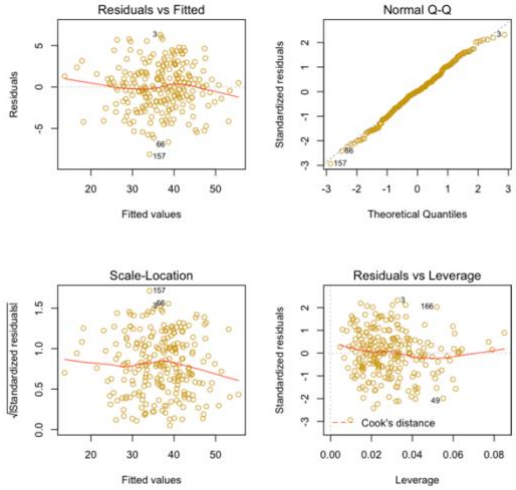

Figure 4: Diagnostics of the linear model evaluated on the testing data.

The residuals in Figure 4 look evenly spread about zero and appear to be normally distributed. There are a couple of outliers but none significant, and there don't appear to be any leverage points looking at the residuals vs. leverage plot. The linear model does indeed satisfy the model assumptions.

One interesting observation to note is that four of the six predictors deemed statistically significant in determining regular season success are statistics against a given team. This could signify that a superior defensive core is arguably more important than a superior offensive core, as it appears preventing the other team from generating offense is more influential than generating offense.

Predicting Game Outcomes

The previous analysis provided insight as to which statistics are most influential in determining a team's overall season success, but what if we want to predict the outcome of a game before it happens? The same 19 statistics were used to do this, but instead of the entire season averages, just the averages over the previous 15 games were used. This allowed for prediction of the outcome of each team's final 67 games but prevented prediction of each team's first 15 games. With team rosters and staff changing so much in the offseason, it's difficult to predict how the team will do before at least a handful of games are played.

To predict the outcomes of the game, the difference of the home and the away teams' 19 statistical averages over the previous 15 games were used as predictors, predicting whether or not the home team would win. Classification techniques of logistic regression, random forests, and support vector machines were all used to train and test the data. With logistic regression, both model selection techniques of backward elimination and AIC were again used.

Classification Method	Accuracy
Logistic Regression	57.68%
Random Forest	54.03%
Support Vector Machines	56.62%

Figure 5: Resulting accuracies from the 3 classification techniques.

It is clear from Figure 5 that the logistic regression model provided the highest accuracy in predicting on the testing dataset at 57.68%. Model selection techniques of backward elimination and AIC with the logistic regression resulted in the exact same model, with three predictors:

$$logit(P(Win)) = -0.054 + 0.19X_{Corsi.d} + 0.16X_{GF.d} - 0.15X_{GA.d}$$

Where $X_{Corsi.d}$ is the difference in average Corsi percentage over the previous 15 games, $X_{GF.d}$ is the difference in average goals for, and $X_{GA.d}$ the difference in average goals against. The difference in Corsi percentage proved to be the most influential at first glance, as a one unit increase in Corsi percentage difference provides a multiplicative odds increase for the home team winning of exp{0.19}=1.21. However, Corsi percentage is in between 0 and 1 so a one-

unit increase is virtually impossible. More notably, a one unit increase in average goals for difference provides a multiplicative odds increase for the home team winning of exp{0.16}=1.17.

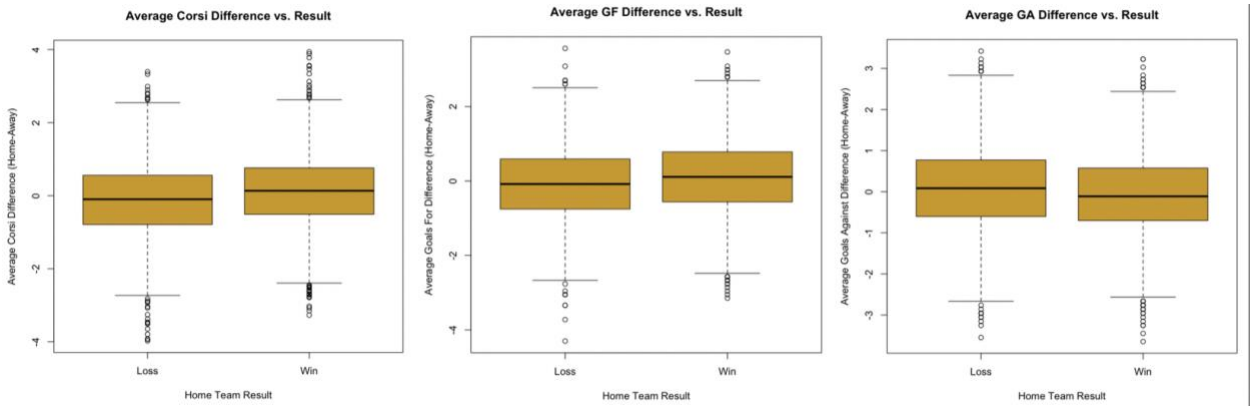

Figure 6: Boxplots of each of the three logistic regression predictors conditioned on home team win/loss.

It's clear from Figure 6 how the three predictors impact the outcome of the game for the home team. A winning home team has a higher average Corsi percentage difference and higher average goals for difference on average while a losing home team typically has a higher average goal against difference

Home.Team	Away.Team	Corsi.Diff	Goals.Against.Diff	Goals.For.Diff	Home.Result	Home.Pred.Results
FLA	WPG	-0.5883315	-2.07414558	-1.7113548	Loss	Loss
WPG	COL	-0.1365746	0.67303700	2.1245197	Win	Win
S.J	DET	0.3855535	0.47680967	-0.1770050	Loss	Loss
NYR	CHI	-0.5433762	0.08435502	1.3573448	Loss	Win
BOS	CBJ	0.6358929	-1.48546360	0.3983762	Win	Win
MIN	BOS	-0.8610017	-0.70055429	-2.6703235	Loss	Loss

Figure 7: Six examples of home team win/loss predictions based on the logistic regression model.

Of the six examples in Figure 7, five of the games are predicted correctly, which is not necessarily indicative of the overall accuracy of 57.68%.

n=1503	Predicted: Loss	Predicted: Win
Actual: Loss	481	363
Actual: Win	273	386

Figure 8: Confusion matrix of the predictions with the logistic regression model.

From the confusion matrix in Figure 8, it's clear the logistic regression model was more accurate in predicting losses than predicting wins. No analysis of the underlying reason for this was done, but it is certainly worth consideration in future models.

Conclusion

The linear regression model proved superior to the binomial regression model in terms of determining an NHL team's regular season success. With an MSE of 5.35, the model predicted ROW of the training set quite accurately, on average being off by 2 ROW. Only three predictors in both analyses were advanced hockey statistics, so perhaps incorporating more of these would improve the accuracy. In predicting single game outcomes, the logistic regression model proved superior with a prediction accuracy of 57.68%. Again, the use of more advanced statistics may allow for even better accuracy with both models. However, hockey odds-makers in Las Vegas tend to predict NHL games at around 60-65%, so there's not a ton of room for improvement, but definitely some. As mentioned previously, it's worth taking a deeper look into why the logistic regression model predicted losses so much more accurately than wins. Fixing this could potentially improve our accuracy a fair amount in future models.