CARLETON UNIVERSITY

SCHOOL OF

MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: **National Hockey League Data:**

Analysis and Prediction

AUTHOR: **Howard Silvey**

SUPERVISOR: **Dr. Jason Nielsen**

DATE: **April 7th, 2020**

Abstract:

"All the statistics in the world can't measure the warmth of a smile" (Hart, 2012).

As a born and raised Canadian, upon nearing completion of an undergraduate degree in statistics, no domain of research seemed more appropriate than the warm-heartfelt game of hockey. This research paper will be studying the game through the lens of the National Hockey League (NHL) and their proprietary data, retrieved through a third-party distributor. The focus of this paper being that of a detailed analysis and prediction model to be crafted for the purpose of future goals scored for any given home team. A multiple regression model with least squares coefficients is constructed from a quantitative response variable Y representing the number of goals scored by the home team any given game. Noting that as the original intention of this project was for use of per-game prediction, a second half of this research project could be carried out with response variable Y representing the visiting team's number of goals per game, then running both analysis' together, or simultaneously as a single model with a vector-valued response (Y_1 , Y_2), to predict future games. Then having built an appropriate model, prediction-testing on both in-and-out-of-sample data can be performed to check the model's predictive power.

A domain of analytics ripe with newfound insights to be discovered, hockey analytics began picking up steam in 2014 when the second Alberta Analytics Conference filled the Calgary Saddledome in attendance for ten different in-depth presentations, aiding what is now referred to as the summer of hockey analytics. This recent prevalence of analytics in hockey and in sports in general can be rooted in the over-encompassing fact that statistical power can drive accuracy well before time can. In other words, the much larger data set offered by tracking various possession and percentages statistics, and using them as predictors for goals, as opposed to simply waiting to observe enough goals over time, is offering the analytics community an avenue for predictive modelling well before the data of just goals or

even shots-on-goal can offer. To date, the three most commonly used statistics in the analysis of ice hockey are “Corsi” and “Fenwick” (both use shots as a proxy to approximate puck possession; shown to correlate well with possession) and "PDO" (considered proxy for random variance; measure of luck), as puck possession (with a dash of luck) tends to be a good predictor of which team has the dominant advantage, making these double proxy measures good game-outcome predictors. This of course is not always the case. Fans of the Toronto Maple Leafs in the 1990’s became all too familiar with strong puck possessions during power-plays in which the puck was cycled in the opponent’s end but was hardly shot on net, ultimately wasting their most advantageous moments. With a plethora of statistics to choose from, puck possession’s strengths and weaknesses will be investigated and best accounted for.

Given the time and resource constraints, the analysis of this project will be team and not player focused. This will drastically reduce the model formation work required, reducing the necessary variable components down from approximately seven-hundred-and-thirteen (31 teams x 23 players in per game roster) to only thirty-one. With every benefit comes a cost. The cost to be paid from such a variable reduction is a rather large one in terms of modelling: accuracy. When creating a per-game predictive model, it would benefit the model best to have each player’s statistics being implemented, to achieve as explicit a model as possible (e.g. model reflects roster, injuries, etc.). On a per team basis (and to a lesser extent a per player basis), the main problem faced when doing an analysis of a sport as dynamic as hockey is multicollinearity; some seemingly independent variables (e.g. player’s contributions) are highly correlated. Luckily, there are techniques at one’s disposal that will help mitigate this problem.

A couple different selection procedures will be used to help ensure the best possible accuracy with regards to the variable selection process. A linear regression model with least squares coefficients will be created using the Forward Selection and Backward Elimination procedures. These procedures will be used over a random sample of the most recent four seasons of NHL data to capture the best predictor variables for the model.

Then, having found a tentative model, precision and multicollinearity can be attacked from the angle of Lasso and Ridge regression analysis. Sadly, due to the time constraint of this research project, both procedures were unable to be utilized. For the interested reader, a more detailed description of these regression procedures is given:

Lasso regression seeks to put constraints on the model in order to have the coefficients on some of the variables shrink toward zero. This allows for a variable selection procedure that is paired with regularization aimed at enhancing the model's prediction accuracy. Having just performed forward and backward variable selection, making use of another variable selection procedure to test the accuracy of the tentative model will add another degree of confidence. Ridge regression hopes to find more reliable predictor estimates by adding a small biasing constant k. The downside to Ridge's regression procedure is it diminishes the strength of the correlations in exchange for reduced multicollinearity, which makes it a good form of regression for testing the collinearity problem upon already choosing a tentative model through other procedures.

Having performed two stepwise regression procedures, predictions of a handful of games occurring just after the sampled data is given. Then comparing, the same number of game predictions is given for out-of-sample data from both early in the 2019 season, and for games in the 2020 season. To then test the predictive power of the model, a Leave-One-Out Cross Validation will be performed; where the model will be tested on the sampled data with one data point removed, for each and every data point. This will help to determine the magnitude of the chosen model's prediction error.

Four seasons of NHL statistics and analytics data will be sifted through to build the model, of which we will use to predict with relative success the outcome of various NHL games. After a best model (given constraints) is found and predictions are made, a comparison to the current state of NHL analytics will be examined. This will position the discussion to explore the possible future of analytics in hockey

and the implications it may have on the game itself. As the game of hockey continues to drive an upsurge in hockey analytics hopefuls, so too will these hopefuls push the long-cherished game of hockey.

Table of Contents:

1 Introduction:

1.1 History of Hockey and Analytics

The history of professionally played hockey can be traced back to the turn of the 20th century. At the time, the National Hockey League was one of the few associations that competed for the Stanley Cup, the oldest trophy of any professional sports league in North America. After superseding the National Hockey Association due to a dispute between team owners, the newly founded National Hockey League begin its journey with humble beginnings of merely five Canadian teams.

Not long after, by 1924 the league saw the introduction of the Boston Bruins, the first team from the United States to join. Only two years later, the league was up to a total of ten teams, of which six were American. The Ottawa Senators saw booming success in the 1920's, with a total of four Stanley Cup wins. The 1930's saw the folding of Ottawa's hometown Senators, and by the 1940's the league was whittled down to six teams, a number it stayed at for over twenty-five years; a period of the NHL's history now coined the Original Six Era.

The end of one era was followed by another, known as the Expansion Period, in which the NHL (out of growing competition with the World Hockey Association) quickly doubled in size to twelve teams, and continued welcoming new teams in bundles for another twenty years. With an increased league size came increased attention to detail, both on and off the ice. New York Rangers center Andy Bathgate stumbled upon the "broken blade" trick used to change the physics of his slapshot, which was not long after picked up by Chicago Black Hawks' forwards Stan Mikita and Bobby Hull. These two teammates then solidified the idea of curved blades for increased shooting accuracy and control by asking their manufacturers to create sticks with pre-curved blades. Decades later, curved blades are now the standard.

Off the ice many different changes were taking place. As the Expansion era was ending, from it sprouted the Goaltender Era, of which the focus on defense and goaltender strength changed many aspects of the game. Goaltenders equipment become strikingly larger, with their stances more strategically lowered to better account for the high proportion of low shots, and coaches better equipped with the knowledge and insight from past game video recordings. As the digital age began to take life, so too did hockey analytics.

With an increased use of these statistics came a better understanding of them. Informally called the worst statistic in hockey, the plus/minus statistic (a player's on-ice goal differential) began being used as early as the 1950's by the Montreal Canadiens. Player's plus/minus statistic use steamrolled across the league until the 1967 season when the NHL began officially keeping track. It started becoming clear to all invested parties that such a simplistic statistic is prone to giving crude results. This is best exemplified by player Paul Ysebaert, who went from having the highest plus/minus in the league at +44 to only six years later having the lowest plus/minus at -43.

With hockey's fluid nature, it is undoubtedly hard to formulate statistics for as it lacks much of the stationarity and independence found in other professional sports such as baseball's batter segments or football's in-play parcels. Though this being the case, over the years continued efforts to overcome the game's complexity have been made. The 1980's brought about a sweeping change in how the NHL recorded its statistics. From handwritten ledgers, the NHL evolved its statistics record keeping to a computer system called the Real Time Scoring System (RTSS), again later upgrading into the Hockey Information Tracking System (HITS). This revolution in data access and accuracy was met with an influx of interested fans and teams alike. Over the recent decades many different statistics have been tracked, and from them new ones created (e.g. Corsi, Fenwick, High-Danger Scoring Chances, etc.). From the

2008-2012 Toronto Maple Leafs' Head Coach Ron Wilson using a laptop on the bench to track statistics such as player's ice time, to more recently in 2014 when the "Corsi" statistic inventor, a financial analyst named Tim Barnes, was hired by the Washington Capitals as an Analytics Consultant.

Now, with virtually every team having its own proprietary data and analytics staff, alongside hockey analytics conferences being held regularly across North America (Carleton University conferences hosted 2015-2019), hockey analytics is rife with discovery to be found at both the professional and amateur levels alike.

1.2 Metric Formation

With a vector-valued response variable Y defined to be the number of goals scored by the home team in any given game, for data structured in per-season format, Y will be defined to be a vector representing the goals scored in that given season. To be scaled into a per-game predictive model, for a model build on the state of each season and the differences between them, the best that can be done is to create a model that appropriately fits the past few seasons, and divide by the eighty-two games played each season. As shot-statistics acts as a double proxy for puck-possession's proxy to goals scored, theoretically the weights from each team on the response variable will be adjusted such that each goal scored is weighted appropriately. This implies that for the predictive modelling of a given team, with properly applied statistical techniques, dividing each team's predicted goals per season by games played said season will result in a properly adjusted per-game response variable.

A small table of seven game prediction estimates is calculated and displayed for both in and out of sample 2019-2020 seasonal data. The in-sample data pertains to games played in all seasons, and is hence divided by 82 games, while two out-of-sample datasets are tested from: the first from the first half of the 2019-2020 season, with the second being that of games from 2020 only.

This project will test and make use of many important predictor variables. Knowingly ignoring player data, predictor variables are chosen out of shooting ("possession"), penalty, goaltender, circumstance statistics, and a few basic game statistics. The predictor variables that will be made use of for this project are:

Shooting Statistics (Corsi/Fenwick proxies for possession):

Shots (S) ; Number of shots by a given team

Shots Against (SA) ; Number of shots against a given team

Shooting Percentage (S%) ; calculated as goals divided by shots on goals

Corsi For (CF) ; All the shot attempts at the net for a given team at even strength, calculated as CF = Shots + Misses + Blocks Against

Corsi Against (CA) ; Shot attempts at the net against a given team at even strength, calculated as CA = Shots Against + Misses Against + Blocks For

Corsi For Percentage (CF%) ; The percentage of all shots that are taken by a team, calculated as CF% = CF/(CF + CA))

Fenwick For (FF) ; Total shots at the net for at even strength except for blocked shots, calculated as FF = Shots For + Misses For

Fenwick For (FA) ; Total shots at the net against at even strength except for blocked shots, calculated as FA = Shots Against + Misses Against

Fenwick For Percentage (FF%) ; Percentage of total Fenwick for a selected team, calculated as FF% = FF/(FF + FA)

Possession Statistics (Expected and Adjusted Goals different proxy):

Expected Goals For (xGF) ; how calculated described below

Expected goals (xG) measures the quality of a shot based on several variables such as assist type, distance and shot angle from goal, whether it was a headed shot and whether it was a big chance. Adding up a team's expected goals can give an indication of how many goals a team should have scored on average, given the shots they have taken

Expected Goals Against (xGA) ; how calculated described above only for against team

Adjusted Goals For (aGF) ; how calculated described below

In order to account for different roster sizes, schedule lengths, and scoring environments, some statistics have been adjusted. All statistics have been adjusted to an 82-game schedule with a maximum roster of 18 skaters and a league average of 6 goals per game and 1.67 assists per goal

Adjusted Goals Against (aGA) ; how calculated described above only for against team

Actual Goals exceeding Expected Goals Differential (axDiff) ; calculated as the difference over expected goals of actual subtract expected goals

Scoring Chances Statistics (Scoring and High-Danger Scoring different proxy):

Scoring Chances For (SCF) ; Count of Scoring Chances for a given team based on a value system*

*Value System – Points are assigned, foremost by region, as diagramed below:

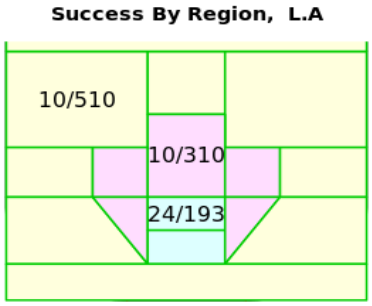


Diagram 1.2: Value System of Scoring Chances For

Where for the colouring of red (center) and yellow (periphery) and green (lining) of an offensive end zone shown above,

Yellow – Add value of 1

Green – Add value of 2

Red – Add value of 3

And secondarily, by circumstance of the shot attempt, such that,

If shot was a *rush* shot (within 4 seconds of puck being in defensive or neutral end of ice without stopping play)

If shot created a *rebound* (any shot attempted within 3 seconds of a blocked shot on net)

Where,

Rush shot – Add value of 1

Rebound shot – Add value of 1

Blocked Shot – Subtract value of 1

Shot made in Neutral Zone – Value of 0 ; excluded (between end zones)
Shot made Defensive Zone – Value of 0 ; excluded

Such that: A shot that results in an end value of 2 is a Scoring Chance.

Scoring Chances Against (SCA) ; how calculated described above only for against team

Scoring Chances For Percentage (SCF%) ; total scoring-chances-for-percentage for a given team, calculated SCF% = SCF/(SCF+SCA)

High Danger Scoring Chances For (HDF) ; scoring chances for a given team with a value of 3 in the Scoring Chances value system*

High Danger Scoring Chances Against (HDA) ; how calculated described above only for against team

High Danger Scoring Chances For Percentage (HDF%) ;

total high-danger-scoring-chances-for percentage for a selected team, calculated as SCF% = SCF/(SCF+SCA)

High Danger Goals For (HDGF) ; Goals Scored from High Danger Scoring Chances

High Danger Goals Against (HDGA) ; Goals Scored Against from High Danger Scoring Chances

High Danger Scoring Chances Percentage (HDC%) ;

Total high-danger-scoring-chances-for-goals-percentage for a selected team, calculated as HDC% = HDGF/(HDGF + HDGA)

High Danger Scoring Chances Opportunities Percentage (HDCO%) ; omitted

Where the remaining statistics are used interchangeable with each class of possession proxy statistics:

Penalty statistics:

Power Play (PP) ; Goals scored during a power play

Power Play Opportunities (PPO) ; Goal opportunities during a power play

Power Play Percentage (PP%) ; Power play percentage, calculated as PP% = PP/PPO

Power Play Against (PPA) ; Power play goals against given team

Power Play Opportunities Against (PPOA) ; Power play opportunities against a team

Penalty Killing Percentage (PK%) ; Percentage of penalties avoided being scored on,

calculated as, PK% = (PPOA* – PP)/PPOA*,

where: *PPOA – Power Play Opportunities Against, is omitted

Goaltender statistics:

Save Percentage (SV%) ; calculated by dividing saves by shots against, SV% = S/SA

Shutouts (SO) ; shutouts (no goals on a given team for the whole game)

Sum of On-Ice Save Percentage (PDO) ; calculation implied in name. Note below.

For PDO, by the law of large numbers, sum percentages ultimately regresses to 100 over time, making it a good proxy for luck as random variance

Circumstance statistics:

Strength of Schedule (SOS) ; a strength of schedule rating system

(system complex – explanation omitted)

Finally, to conclude defining statistics the well-known **Basic Game Statistics** are noted:

Wins (W) ; Number of wins for a team each season

Losses (L) ; Number of losses for a team each season

Points (PTS) ; Number of points for a team each season

Points Percentage (PTS%) ; Percentage of points for a team each season, calculated as

$$PTS\% = PTS / \max(\text{Points Possible})$$

Overtime Losses (OL) ; Losses for a given team resulting from overtime

Average Age (AvAge) ; Average Age for a given team per season

2 Data Organization:

2.1 Hockey's Harold

The data used for this research project was extracted off Hockey-Reference.com, who's primary data provider is Sportradar. Since 2015 the one of the leading data intelligence providers, Sportradar, has had an exclusive third-party partnership with the National Hockey League to handle and distribute the leagues patented data.

All league and per season data were mined from Hockey-Refence.com before the start of the 2020 New Year, excluding the out-of-sample 2020 prediction sample. Any implementation of the model(s) found in this project should be subject to an updated 2019-2020 season for better accuracy if to be applied to further seasons.

Hockey-Reference.com supplies updated data on all components of the game, from player, team, and season data, to more frivolous aspects such as player's birthdays and twitter account information. Once a selection is made, various statistics and analytics tables appear that can be viewed or extracted. Of the many ways offered to extract or share the site's data, the most convenient for use with the programming language R was downloading the table as a Comma Delimited File (*.csv).

Both statistics and analytics data from the 2016-2017 to 2019-2020 seasons were made use of. Data vectors were created out of each file's column of observations, where for the 40 different

preliminary variables tried from, each utilize: 1 season x 30 teams + 3 seasons x 31 teams = 123 rows of team data.

This gives for a total of 40 x 123 = 4920 data observations used in establishing a tentative model. In building such a model many of the variables are found to be either statistically insignificant (variance not explained by response variable) or highly correlated, and therefore are quickly discarded in the variable selection process.

3 Model Design:

3.1 Multiple Linear Regression

The purpose of a multiple linear regression model is to attempt to explain the relationship of a dependent response variable Y and a set of independent predictor variables $X_1,\ldots,Xp$, with $p \geq 2$. Then making use of the three predictive sample sets, predictions are made from the samples of size n=123. Thus, for a sample size of $n \geq 30$, the Central Limit Theorem can be applied. Let n the number of observations of the response variables Y, then the distribution of random samples of size n, taken from a population (with replacement) having mean μ and standard deviation σ, normalizes (regardless of the initial population distribution). This gives an in-sample mean of $\hat{\mu}$ and standard deviation $\hat{\sigma}$:

$\hat{\mu}$ = {Sample mean of goals per game from chosen seasons}

$$= \frac{(2.73) + (2.97) + (3.01) + (3.02)}{4}$$

= 2.9325 , where 2.73, 2.97, 3.01, and 3.02 are the league average goals per game for the 2016-2017 to 2019-2020 seasons respectively

$$\hat{\sigma} = \frac{\{\text{Sample standard deviation of goals per game}\}}{\sqrt{n}} = \hat{\sigma}_P/\sqrt{n} = \hat{\sigma}_P/(\sqrt{123})$$

With a properly defined response variable for the set (X_1,Y_1), (X_2,Y_2),…,(X_n,Y_n) of n observations, the multiple linear regression model is expressed as:

$Y_i = \beta_0 + \beta 1 X_{i1} + \beta 2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$, for all $i = 1,\ldots,n$,

where ε_i are the deviations of the observed values Y from their means $\bar{Y}$.

Referencing heavily the *Applied Regression Analysis* textbook (Draper, N. R., et al., 1981) assigned as Carleton's introduction to regression analysis class (Spring 2018), following the Least Squares Estimation section, it is known that the error term vector has a mean of zero with uncorrelated elements within, namely $E(\varepsilon)=0$ and $V(\varepsilon)=\mathbf{I}\sigma^2$, where **I** is identity matrix with n dimensions.

Thus, implying a response variable's mean of $E(Y) = X\beta$, where Y is known to be a nx1 dimensional vector and β a (p+1)x1 dimensional vector. The parameters β_j for j=0,…,p for the independent variables $X_{i1,\ldots,} X_{ip}$ are therefore estimated by the least square estimates $\hat{\beta}_j$.

Having defined the estimates, least squares estimation for a multiple linear regression operates by minimizing the sum of squares of deviation of the error from the true values (or concisely the error sum of squares). Then for parameters β_j in a multiple linear regression model estimated by least squares estimates $\hat{\beta}_j$, the least squares estimation process for β_j is done by minimizing the sum of squares of errors:

$$S = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2$$

for the variable β_j by setting $\partial S/\partial \beta_j = 0$ and solving for β_j.

This will result in p+1 equations representing p+1 unknown β_j's, which by solving the system of equations as described above produces the estimates $\hat{\beta}_j$. All estimates and their relative importance with be displayed later in tabular form in the analysis section, after selection procedures and the cross-validation are formally defined. Defining the error sum of squares in matrix form, further employing *Applied Regression Analysis,*

$$\varepsilon'\varepsilon = (Y - \beta X)'(Y - \beta X),$$

where the estimates $\hat{\beta}_j$ are found such that in belonging to the (p+1)x1 dimensional vector $\hat{\beta}$ they minimize $\varepsilon'\varepsilon$ for the parameter β. Again, due to the Central Limit Theorem, this process normalizes for a sufficiently large random sample of n ≥ 30. The normal equations for this process are defined by,

$$(X'X)\hat{\beta} = X'Y,$$

where for in the non-singular matrix $(X'X)$ (implying independence of p+1 equations),

$\hat{\beta} = (X'X)^{-1}X'Y$ is the Least Squares Estimator that minimizes the normal equations.

Having defined the error term ε and the parameter β, three important properties of the estimator $\hat{\beta}$ are stated:

(1) The distribution of the error sum of squares $\varepsilon'\varepsilon$ is irrelevant in the minimization process due to the Central Limit Theorem.

(2) Property (1) implies a minimum variance is always achieved for the parameter β with unbiased estimate $\hat{\beta}$, which is comprised of (p+1)x1 linear functions of the observations $Y_1, \ldots, Y_n$.

(3) For normally distributed independent errors such that $\varepsilon \sim N(0, \boldsymbol{I}\sigma^2)$, the estimate $\hat{\beta}$ is the Maximum Likelihood Estimate of the parameter β.

Due to such powerful properties, Least Squares is a simple yet robust methodology that gives it the advantage as a good introductory tool for such a hearty domain of analysis.

Before turning to the selection procedure, a formal reminder of the simple statistic employed to assess the quality of the multiple linear regression is stated:

R^2 - The variation of the model expressed as a percentage ranging from 0 to 1, or,

stated more explicitly: The proportion of the variance in the response variable that is predictable from the independent variable

(also called the *Coefficient of Determination)*

Where the statistic is defined:

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

With,

$\sum(\hat{y}_i - \bar{y})^2$ being the "regression sum of squares" which measures how far the sloped regression line, $\hat{y}_i$, is from the horizontal sample mean, $\bar{y}$

And,

$\sum(y_i - \hat{y}_i)^2$ the "total sum of squares" quantifies how much the data points, y_i, vary around their mean, $\bar{y}$

Thus, the R^2 is interpreted as explained variation (regression sum of squares) divided by the total variation (total sum of squares). With this tool an assessment of the quality of the model's fit will be made in the analysis section of this report.

3.2 Variable Selection Procedure

Following the notations used in *Applied Regression Analysis,* for a set of predictor variables

$X_1,\ldots, X_p$, there exists the set $Z_1,\ldots, Z_r$ of Least Squares functions of any linear, square, or inverse relation in order to accomplish the best fitted model. This can be done through selection of a set up to an index i, where for the best-established subset to said index that minimizes the bias, a trade-off of an increasing variance to the estimates and their reliability is found (known as the *Bias-variance* trade-off).

For the Least Squares estimates $\hat{\beta}$ a balance is stuck by choosing the set $Z_1,\ldots, Z_t$ up to some average bias-variance at observation t, as this will allow for easy calculation of the variance of the estimator of observation t, for the response variable $\hat{Y}_{t,}$ which is $\sigma^2_t = t\cdot\sigma^2/n$. The stepwise regression procedure will be utilized in handling this trade-off. The stepwise regression procedure is an appropriate choice of the many variable selection procedures available as, for starters, it is easily applicable to an initially low number of variables (hence the "step" is stepwise). Additionally, this selection procedure evaluates the weights of the variables as they are assessed, signifying their significance as the process develops. Stepwise regression has two component procedures to work with, Forward Selection and Backward Elimination, both of which are defined formally in detail in the following sections. Regardless of the procedure used in Stepwise Regression, both the Forward Selection and Backward Elimination procedures choose the estimates with the highest predictive power by checking two different significant thresholds. As is standard practice, a P-value level of significance less than or equal to 0.05 (5%) is first used, with the verification step utilizing partial F-values after already identifying the most or least significant variable (depending on the procedure used). The desired partial F-value of a set $Z_1,\ldots, Z_i$, for some Z_i added to the current state of the regression model, the partial F-values are calculated as, for coefficient $\hat{\beta}_i$,

$F_i = (\hat{\beta}_i / s.e(\hat{\beta}_i))^2 \sim F(1, n-\upsilon, 1-\alpha)$, where υ is the number of predictors in the model

before the addition of Z_i , and α the significance level.

Knowing the distributional properties of the F-distribution, the F-distribution as calculated above is equivalent to the square of the t-distribution for a sample size of n, implying the use of t-values in the coming analysis. Thus, upon checking a predictor through the use of the P-value threshold, confirming the given t-value that has the largest or smallest absolute t-value depending on the procedure will verify or refute said predictor's significance.

3.2.1 Forward Selection

The procedure for Forward Selection begins with as rudimentary a model as possible (excluding the trivial $Y = \beta_0$ intercept model), namely,

$Y = \beta_0 + \varepsilon$, with the model's intercept β_0 and vector of errors ε.

Then, the predictor variable Z_j that has the largest correlation with the response variable Y is added to the regression equation, giving,

$Y = \beta_0 + \beta_j Z_j + \varepsilon$, for parameter β_j .

Proceeding in this manner, the next predictor Z is found through the use of the dual threshold procedure, whereas the Forward Selection procedure seeks out the predictor with the lowest P-value to add, thus implying that a largest t-value threshold shows significance. Once that specific Z_h within the set of predictors is chosen and added to the model, the R^2 statistic is checked to assess the quality of the model's fit against that of its previous state prior to the addition of the last predictor. Continuing in this fashion, the predictor variables P and t-values are compared at each iteration and the model is updated or unchanged to best reflect the strongest correlation with the dependent variable Y. Not forgetting to note, all predictor variables in the model must be re-evaluated upon addition of a new predictor, as the

relative significance of each variable may change due to the newfound affecting correlation. Therefore, after adding all significant predictor variables with removal of insignificant ones at each step, such that the variable with the highest significance not in the model is insignificant, a best model by the Forward Selection procedure is achieved.

3.2.2 Backward Elimination

Moving to Stepwise Regression's procedural second half, Backward Elimination proceeds as the name implies: starting with a general predictive model of all variables to be tested from, Z_1,…, Z_r, the multiple linear regression model takes the form,

$$Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_r Z_r + \varepsilon \text{, for a vector of errors } \varepsilon.$$

The goal of this essential second part is to acquire the same regression model as was found with the Forward Selection process. This may not always be true, as it is the case that either process is subject to a creating misleading model, due to a handful of reasons such as overfitting, underfitting, or simply an inadequate analytical toolkit. Just as is the case with the first procedure, only reversed, a threshold significance level of a P-value ≤ 0.05 is used in the elimination process. Backward Elimination proceeding such that the predictor variable with the highest P-value is removed at each step, whereby in checking the t-values, the smallest absolute t-value confirms 'least significance'. Also, remembering to have the model have its R^2 statistic checked and recorded at each iteration of the removal process. Unlike with Forward Selection, the set of the variables' correlation strengths with regards to one another is known by starting with all the predictor variables, thus implying no need to check the other variables after the smallest t-valued predictor is removed. At some step, when the model finds itself with only predictors of significant P-values (P ≤ 5%) and (hopefully) has an R^2 statistic of reasonably strength, a best model by

Backward Elimination is established.

If performed properly, the Backward Elimination procedure should result in the same predictive model as Forward Elimination, adding validity to the model's predictive power. Still, as it is common that the procedures converge to the same model only to come across more general problems such as overfitting, a validation procedure is formally defined to further build on the model's validity.

3.3 Leave-One-Out Cross-Validation

In evaluating the appropriateness of the chosen model, a Leave-One-Out Cross-Validation is applied. Stated generally, a cross-validation procedure attempts to test the validity of the models fit through the means of training and testing sets created out of testing data (as opposed to checking model validity through residual values). This form of validation can help give the model an extra degree of predictive power as it creates a more robust model for out-of-sample data, where the general goal of most models being that it be applicable to independent data sets.

The process begins by taking n observations of the data and dividing them into training and testing sets, excluding the use of a validation set given the aforementioned time constraint diminishes its return. As is expected with an initial testing set of 40 predictor variables, the model being built out of a large array of statistics to choose from is subjugated to potential overfitting. Overfitting, in terms of the signal that the model attempts to distinguish, and the noise that surrounds that signal, can be defined and the over-valuing of the noise in the data under study. Regarding this research project, this implies that the created model would work to a fair degree on the 2016-2019 seasonal data but would have a significantly noticeable drop in predictive power for all other seasonal data. To test for overfitting, the model will be trained on one set of data and tested for overfitting on another.

The type of cross-validation being utilized is a Leave-One-Out Cross-Validation, defined to be a

type of K-Fold Cross-Validation evaluation procedure in which a testing set of size 1 is trained on a set of size n-1. This is done in such a way as to train all of the data in the training set on n-1 data points, whereby a prediction is created on each excluded point by running n separate tests. As implied by the name, a K-Fold Validation procedure splits the data into k separate training sets, a step up from the simplest cross-validation technique the Holdout Method (In which the data is split into only a training and testing set). A Leave-One-Out Cross-Validation takes the K-Fold Validation to the extreme, splitting the data between a minimized testing set and a training set of n-1 over n iterations.

While overfitting may be attacked from the angle of the Leave-One-Out Cross-Validation, the externality presented from using such a method happens to be the price of the test error itself. This is because as the predicted data point is only run on a testing set of size 1, implying that the error between the actual response variable (Y) and the predicted response variable ($\hat{Y}$) may be quite large. Luckily, as detailed above in its definition, the Leave-One-Out Cross-Validation process runs n iterations of the validation procedure, allowing for a comparison of the n testing data points against the n-1 training set data points. Thus, the Mean Squared Error (MSE) for a given set of n observations, i = 1,…,n, having been relatively high in one iteration: $MSE_i = (y_i - \hat{y}_i)^2$, shrinks in size over n iterations:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

Making MSE a good estimate in testing the error of the model.

While many other forms of model validation exist such as the Set Approach or Game Simulations, a Leave-one-Out Cross-Validation approach is taken as the MSE estimate stays consistent due to the reproducible nature of the validation procedure. This nature is the result of the process having a methodical (entirely non-random) training set selection process, which as a result

systematically reproduces the same MSE test error over n iterations.

Now, having established the groundwork in describing and defining the different statistical tools and concepts to be used in this project, an analysis of the sample data may begin.

Note: For predictor variables with longer than usual names an acronym is given instead, followed by a "*" indicating the need to refer to *1.2 Metric Formation* for the appropriate variable name.

4 Analysis:

4.1 Model Selection

The analysis begins with the Backward Elimination procedure. Starting with all predictor variables and their estimates, the predictor variables FA and OL produced NA's (Not available) and hence were removed without affecting any of the other estimates. Displaying the first summary of 38 predictors:

	Estimate	Std.Error	t-value	Pr(>|t|)	
(Intercept)	637.44145	2453.87118	0.260	0.7957	
S	0.01091	0.12517	0.087	0.9307	
SA	0.06868	0.11852	0.579	0.5638	
S_Percent	-3.32198	25.65250	-0.129	0.8973	
CF	0.21442	0.13044	1.644	0.1039	
CA	-0.21768	0.12175	-1.788	0.0774	.
CF_Percent	-29.78884	16.67732	-1.786	0.0776	.
FF	-0.10163	0.15005	-0.677	0.5001	
FF_Percent	15.26208	13.66393	1.117	0.2672	
xGF	0.11772	1.35704	0.087	0.9311	
xGA	-1.56616	1.52832	-1.025	0.3084	
aGF	0.34217	1.43896	0.238	0.8126	
aGA	0.14779	1.32613	0.111	0.9115	
axDiff	-0.08973	0.36472	-0.246	0.8063	

SCF	-0.03486	0.31047	-0.112	0.9109
SCA	0.12285	0.32290	0.380	0.7045
SCF_Percent	-1.35138	17.66951	-0.076	0.9392
HDF	-0.54656	0.50360	-1.085	0.2809
HDA	0.68806	0.58154	1.183	0.2400
HDF_Percent	4.81625	5.11875	0.941	0.3494
HDGF	-0.23042	1.26988	-0.181	0.8564
HDGA	0.38397	2.13468	0.180	0.8577
HDC_Percent	1.06945	4.86971	0.220	0.8267
HDCO_Percent	1.76909	8.77618	0.202	0.8407
PP	-1.10412	3.43730	-0.321	0.7488
PPO	-0.03151	0.79213	-0.040	0.9684
PP_Percent	7.53615	7.88946	0.955	0.3422
PPA	-1.58254	1.99650	-0.793	0.4302
PPOA	0.10176	0.40495	0.251	0.8022
PK_Percent	-2.19984	4.92945	-0.446	0.6565
SV_Percent	108.98038	2535.47510	0.043	0.9658
SO	-0.28128	2.94341	-0.096	0.9241
PDO	0.02743	24.92170	**0.001**	**0.9991**
SOS	161.73090	167.78450	0.964	0.3378
W	2.41883	6.50378	0.372	0.7109
L	3.64155	4.11330	0.885	0.3785
PTS	-0.14874	4.37766	-0.034	0.9730
AvAge	0.63701	5.66797	0.112	0.9108

R^2: 0.2145 / adjusted R^2: -0.1275 **|** Observations: 4920

R-Console Printout 4.1: Backward Elimination Printout

It is clear that PDO displays the highest P-value and smallest absolute t-value, and thus can be removed. Omitting all further iterations, the Backward Elimination process proceeds for 32 more steps, until only the predictors CF, CA, CF_Percent, PP_Percent, and L remain. This is done with both P and t-

value thresholds of significance, wherein this best-found model using Backward Elimination consisting of these 5 variables has a Coefficient of Determination of $R^2 = 0.1161$ (11.61%).

Now proceeding to the Forward Selection process, the first predictor is found by way of the highest correlation with the response variable, which can be found by the t-value and P-value columns of each variable's regression summary:

*R^2 values checked at each step (omitted)

	Estimate	Std.Error	t-value	Pr(>|t|)	
(Intercept)	235.527151	21.197752	11.111	<2e-16	***
S	-0.009519	0.009033	-1.054	0.294	
(Intercept)	242.419677	20.974400	11.558	<2e-16	***
SA	-0.012516	0.008935	-1.401	0.164	
(Intercept)	194.758	40.207	4.844	3.81e-06	***
S_Percent	2.026	4.291	0.472	0.638	
(Intercept)	231.871652	16.178182	14.332	<2e-16	***
CF	-0.005613	0.004807	-1.168	0.245	
(Intercept)	236.687412	16.228206	14.585	<2e-16	***
CA	-0.007095	0.004824	-1.471	0.144	
(Intercept)	179.0509	89.3069	2.005	0.0472	*
CF_Percent	0.6919	1.7847	0.388	0.6989	
(Intercept)	231.726237	16.168256	14.332	<2e-16	***
FF	-0.007481	0.006455	-1.159	0.249	
(Intercept)	162.137	90.459	1.792	0.0756	.
FF_Percent	1.030	1.807	0.570	0.5698	
(Intercept)	238.0314	18.0570	13.182	<2e-16	***
xGF	-0.1583	0.1139	-1.389	0.167	
(Intercept)	239.0527	18.1575	13.166	<2e-16	***
xGA	-0.1649	0.1146	-1.439	0.153	

```
(Intercept) 236.9043    19.2702  12.294   <2e-16 ***
aGF          -0.1633     0.1320  -1.237    0.218


(Intercept) 224.34523   19.66069  11.411   <2e-16 ***
aGA          -0.07516    0.13473  -0.558    0.578


(Intercept) 214.725614   4.892424  43.889   <2e-16 ***
axDiff       -0.001653   0.003715  -0.445    0.657


(Intercept) 235.96890   16.91039  13.954   <2e-16 ***
SCF          -0.01546    0.01134  -1.364    0.175


(Intercept) 237.39494   17.03982  13.932   <2e-16 ***
SCA          -0.01645    0.01143  -1.439    0.153


(Intercept) 2.133e+02  8.024e+01   2.658  0.00893 **
SCF_Percent 7.531e-03  1.603e+00   0.005  0.99626


(Intercept) 231.18120   12.73327   18.16   <2e-16 ***
HDF          -0.05790    0.03966   -1.46    0.147


(Intercept) 228.05787   12.83088   17.77   <2e-16 ***
HDA          -0.04759    0.03999   -1.19    0.236


(Intercept) 243.6055    52.6028   4.631 9.23e-06 ***
HDF_Percent  -0.5992     1.0483  -0.572    0.569


(Intercept) 230.1529    11.7904  19.520   <2e-16 ***
HDGF         -0.3320     0.2215  -1.499    0.136


(Intercept) 218.7012    12.1457  18.006   <2e-16 ***
HDGA         -0.1019     0.2288  -0.445    0.657


(Intercept) 209.3690    15.1657  13.805   <2e-16 ***
HDC_Percent   0.3473     1.1858   0.293     0.77



(Intercept)   164.119     30.888   5.313 4.98e-07 ***
HDCO_Percent    3.522      2.177   1.618    0.108


(Intercept) 205.2718    15.9321  12.884   <2e-16 ***
PP            0.1922     0.3530   0.544    0.587
```

```
(Intercept) 231.83748   21.08318  10.996   <2e-16 **
PPO          -0.08264    0.09377  -0.881    0.38


(Intercept)  165.940     24.053   6.899 2.57e-10 ***
PP_Percent     2.417      1.201   2.013   0.0463 *


(Intercept) 232.9624    17.4620  13.341   <2e-16 ***
PPA          -0.4442     0.3895  -1.141    0.256


(Intercept) 215.182479   5.366901  40.094   <2e-16 ***
PPOA        -0.001893   0.004036  -0.469    0.64


(Intercept) 191.5124   110.8204   1.728   0.0865 .
PK_Percent    0.2759     1.3812   0.200   0.8420


(Intercept)    875.1      451.8   1.937   0.0551 .
SV_Percent    -729.5      498.2  -1.464   0.1457


(Intercept)  223.700      8.145  27.466   <2e-16 ***
SO            -2.332      1.616  -1.443    0.152


(Intercept)  637.027    370.056   1.721   0.0877 .
PDO           -4.234      3.700  -1.144   0.2548


(Intercept)  213.623      4.235  50.446   <2e-16 ***
SOS           34.676    109.191   0.318    0.751


(Intercept) 229.4723    15.9536  14.384   <2e-16 ***
W            -0.4304     0.4181  -1.029    0.305


(Intercept) 220.2251    14.5064  15.181   <2e-16 ***
L            -0.2321     0.4886  -0.475    0.636


(Intercept) 233.1465    17.0837  13.647   <2e-16 ***
PTS          -0.2380     0.2019  -1.179    0.241


(Intercept)  248.224    123.577   2.009   0.0468 *
AvAge         -1.239      4.424  -0.280   0.7799
```

R-Console Printout 4.2: Forward Selection Printout

As can been seen above, PP_Percent is best correlated with Y through means of a smallest P-value and largest t-value and is therefore the first variable to be added to the regression model. Verifying once that the t-value relation holds, checking the F-Ratio against its critical F-value using ANOVA, it is found that with an F-Ratio of 4.053 on 1 and 121 degrees of freedom in the numerator and denominator respectively, for a found critical F-value of 1.77 at the 0.05 significance level by checking an F-table, there is a rejection of the null hypothesis and PP_Percent is added to the model.

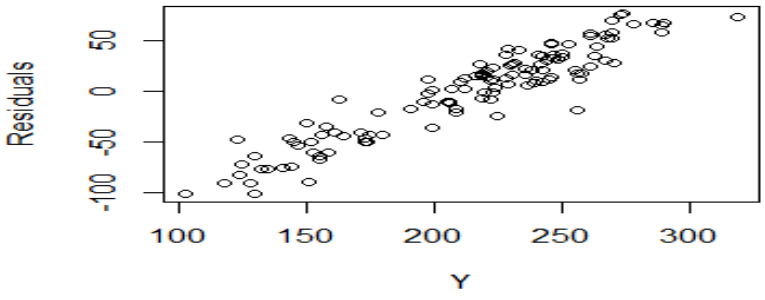
After removing the NA variables (FA and OL) without affect to any estimates, a second step is undergone, and it is found that PDO has the best correlation with P-value ≤ 0.04465 passing the 5% significance threshold, and the absolute t-value=-2.029 being the largest, and is thus added to the model. Noting that on this second step, and all thereafter, the correlation strength of all previously added variables, that being PP_Percent here, is tested for significance with the response after the addition of PDO. It is found that the significance of PP_Percent is still well within the desired threshold; therefore it remains in the model. Also noting the R^2 statistic increased slightly, indicating the model is slightly more predictive with the addition of PDO.

Pursuing the above procedures, the next step for the third sought variable is performed, and no additional variables are found, at neither the required p-value nor verifying t-value level of significance. This implies that the best-found model using Backward Elimination is composed of only the variables PP_Percent and PDO.

As stepwise regression has its many limitation, especially when applied to large multicollinear datasets, different resulting models from each procedure is to be expected. Digging further into each model, as the Backward Elimination model has a smaller residual error on less degrees of freedom, at first glance it appears to be a better model. Then noting Backward Elimination's five predictor variables

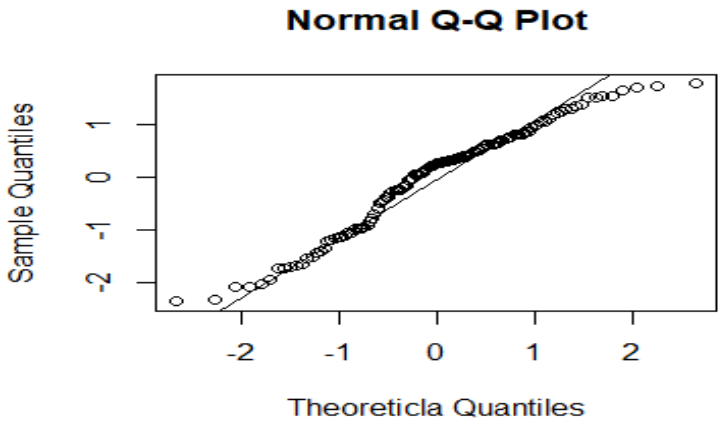
over forward's three, which is to be viewed as advantageous as too few variables can lead to biased estimates. It is also worth noting that both models were the outcome of varying tests of significance, at both initial check of a p-value≤ 0.05 and p-value≤ 0.10. This was done in attempt to minimize human error. In regression modelling, a trade-off can be found in including less significant variables in exchange for increased predictive performance of the model. Thus, to follow statistical procedural techniques as best aligned to an undergraduate student's understanding as possible, both of the commonly used significance levels of 5% and 10% were applied (some papers cite 15-20% level of significance for predictive purposes). No converging nor changed model was found at either significance level, giving greater weight to Backward Elimination's model being the better choice. Finally, noticing that the backward procedure's model has a better predictive performance, with a higher R^2 (and even adjusted R^2 which shrinks stepwise overestimation) of $R^2 = 0.1161$ over Forward Selection's R^2 statistic of R^2=0.06451. Although such a low R^2 is not ideal for actual prediction purposes, given this project's constraints, it is determined to be the "best" found model.

Moving onto the residual analysis, where the residual is formally defined as the deviation of the dependent variable from the fitted values (Y - $\hat{Y}$), a first glance at the plot of the residuals is taken:



Graph 4.1: Scatterplot of Residuals vs. Response Variable

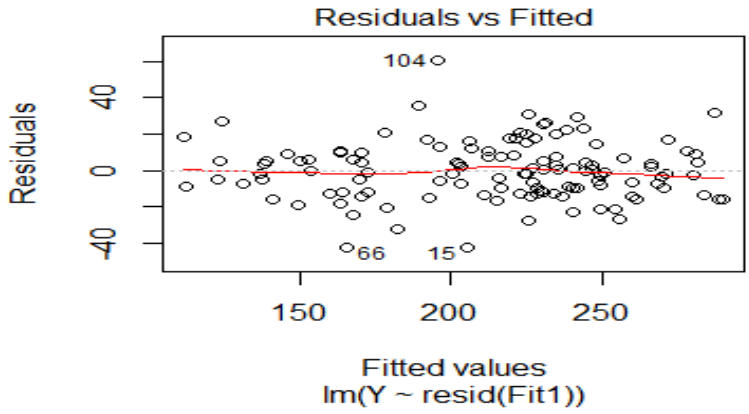
It is easy to see that the errors appear to be correlated, as uncorrelated errors tend to scatter randomly about the mean. In looking for systematic trends affecting the predictive power of the model, the QQ-Normality plot, which tests if the error terms are normally distributed or not, is utilized and displayed below:



Graph 4.2: QQ-Normality Plot Using Residuals

As can be seen, the values appear fairly normal with only a few outliers amidst the upper quartiles. Thus, for nearly two quantiles from the mean in either direction, the residual values appear fairly normal about the mean.

Moving to a more dynamic understanding of the residuals, next the plot of residuals against its fitted values (estimated responses) is viewed below:

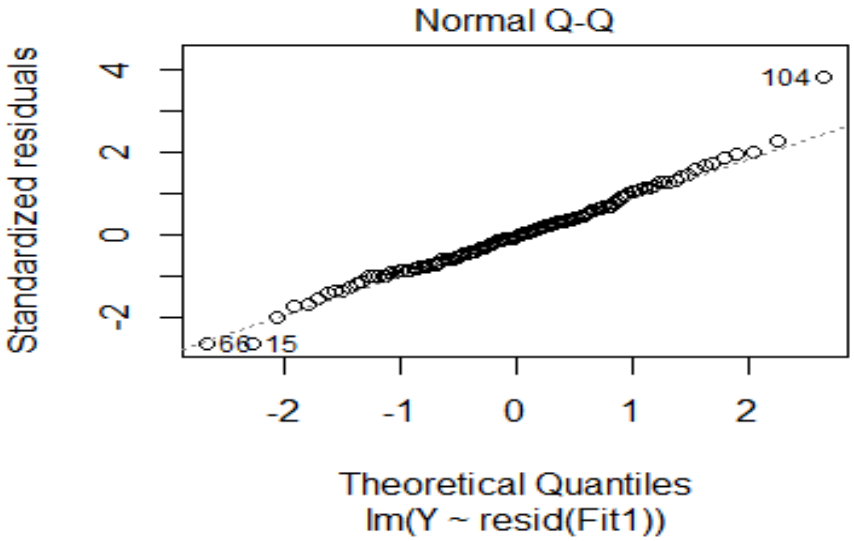


Graph 4.3: Residuals vs. Fitted Values Plot

As can be seen above, the line of best fit oscillates around a mean centred about zero and the data points appear relatively scattered with no discernible pattern in the data. This is indication of a good residuals vs. fitted values plot, or more precisely, an indication of a linear relationship between the variables (linearity) and a relatively equal variance along the regression line (homoscedasticity).

Note: A quadratic variable was tested for each variable and was found not to improve the model's fit (by it via variable or error significance).

Another promising discovery is the well normalized standard residuals against fitted values QQ-Normality plot:



Graph 4.4: QQ-Normality Plot Using Standardized Residuals

Showing through the standardized residuals that the residuals hold normality. This is a necessary condition in conducting an accurate model for inferencing (e.g. confidence intervals, predictions) as it promises that the distribution of the error's subgroups is the same. Broadly speaking, this implies that the errors created by the model will be consistent across variables and observations (i.e. they are random errors).

Now having verified random errors for a relatively suitable model, the final step of the analysis can begin; model assessment.

4.2 Model Assessment

Having weaved through the selection procedure and established a relatively good model alongside a successful residual analysis, a model assessment can be performed. The assessment begins with predictions of in-and-out-of-sample data. This is followed by a Leave-One-Out Cross-Validation

to help establish the limitations of the model by calculating the prediction test estimates of the model at each iteration of the leave-one-data-point-out process. The total prediction error, which takes the average of all of the test error estimates, is represented by the Root Mean Squared Error (RMSE), and the value of the R^2 statistic under the validation process, displayed alongside.

The first few predictions and their corresponding real values (home team goal for) are given below, recalling that data points are on a per-season basis and need be divided by total games per season to get an accurate reading:

*Confidence intervals information omitted

Table 4.1: First Seven Sample Data Predictions (To 2 decimals)

		Per-Game Home Team Goals For	
Date	*Game*	*Prediction*	*Real Home Team # Goals*
01/09/2020	Kings at **Golden Knights**	210.5839/82 = **2.57**	**2**
01/10/2020	Coyotes at **Hurricanes**	204.9252/82 = **2.50**	**3**
01/10/2020	Penguins at **Avalanche**	189.3074/82 = **2.31**	**3**
01/10/2020	Senators at **Red Wings**	198.6757/82 = **2.42**	**3**
01/11/2020	Canucks at **Sabres**	200.9571/82 = **2.45**	**3**
01/11/2020	Kings at **Hurricanes**	192.1530/82 = **2.34**	**2**
01/11/2020	Oilers at **Flames**	214.3782/82 = **2.61**	**4**

Which as can be seen, the next 7 games outside of the sample appear to predict the home teams' goals for with a fair degree of accuracy. All games tested fell within 1.5 goals from the desired value (2 even rounding to the correct value).

Now displaying estimates of the next seven games for the model tested on out-of-sample data from earlier in the 2019 season. As these predictions apply new data, the model's fit to the given data is represented by the R^2/Adjusted R^2 given below the table:

*Confidence intervals information omitted

Table 4.2: First Seven Out-of-Sample 2019-Season Predictions (To 2 decimals)

		Per-Game Home Team Goals For	
Date	*Game*	*Prediction*	*Real Home Team # Goals*
10/16/2019	Flyers at **Oilers**	**3.84**	**6**
10/16/2019	Avalanche at **Penguins**	**3.82**	**3**
10/16/2019	Hurricanes at **Sharks**	**3.48**	**5**
10/16/2019	Maple Leafs at **Capitals**	**4.06**	**4**
10/17/2019	Predators at **Coyotes**	**3.59**	**5**
10/17/2019	Lightning at **Bruins**	**3.40**	**3**
10/17/2019	Red Wings at **Flames**	**3.73**	**5**

Observations: 4920

R^2 / Adjusted R^2 : 0.0716 / 0.0319

Which shows the predictor variables to have a weaker predictive performance with this out-of-sample data, with the R^2 dropping down to 0.07 versus the in-sample's R^2 of 0.12. Also, it can be seen that the prediction estimates perform worse than before, though with two estimates again rounding to the correct value.

Now displaying estimates of the next seven games for the model tested on out-of-sample data from later in the 2020 season. As these predictions apply new data, the model's fit to the given data is represented by the R^2/Adjusted R^2 given below the table:

*Confidence intervals information omitted

Table 4.3: First Seven Out-of-Sample 2020-Season Predictions (To 2 decimals)

		Per-Game Home Team Goals For	
Date	*Game*	*Prediction*	*Real Home Team # Goals*
01/30/2020	Canadiens at **Sabres**	**2.85**	**1**
01/30/2020	Predators at **Devils**	**2.82**	**5**
01/30/2020	Lightning at **Ducks**	**2.81**	**3**
01/31/2020	Golden Knights at **Hurricanes**	**2.42**	**3**
01/31/2020	Blues at **Oilers**	**2.45**	**4**
01/31/2020	Red Wings at **Rangers**	**2.34**	**4**
01/31/2020	Capitals at **Senators**	**2.61**	**3**

Observations: 4920

R^2 / Adjusted R^2 : 0.0458/ 0.0050

As expected, as this data was from later 2020 season games, and the model was created on data up to the end of 2019, this prediction set has the weakest predictive performance yet, with the R^2 dropping down to 0.046. Also, it can be seen that the prediction estimates perform roughly as bad as before, though again two estimates round to the correct value.

The weakened performance of both out-of-sample datasets was to be expected as it is almost surely a result of the model overfitting to the original data. A validation procedure through the means of a Leave-One-Out Cross-Validation is executed as outlined in Section 3.3.

Utilizing the "Caret" library in the programming language R, a snapshot of the Root MSE (RMSE) and the R^2 output values from the "train()" functioning is given below:

```
Linear Regression

123 samples
  5 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 122, 122, 122, 122, 122, 122, ...
Resampling results:

  RMSE      Rsquared    MAE
  46.20022  0.03927991  37.91543

Tuning parameter 'intercept' was held constant at a value of TRUE
> |
```

R-Console Printout 4.3: Leave-One-Out Cross-Validation Values

Noting that large RMSE values greater than or equal 0.5 implies a model's poor ability to accurately reflects the predicted values, especially paired with a low R^2 value (generally want greater

than 0.6 or 0.7). Given the extremely larger RMSE of 46.20 and an R^2 equal to 0.039, the validation procedure has demonstrated the prediction error to be quite large. Thus, as was to be expected, the best-found linear model performed poorly in accurately predicting goals for a sport as dynamic as ice hockey.

5 Conclusion:

5.1 Summary of Findings

Although other regression models and variable selection procedures should be tested if more reliable prediction estimates are desired, some insights into the analysis of the game itself can be had from the limited model produced in this paper.

Noticing from each prediction sample, as the model was built on a per-season basis and adjusted for predictions of each game, the prediction estimates tend to appear fairly constant over certain clusters. Thus, considering this, a model built on each individual game of the season with each team's statistic would be able to capture the game-to-game fluctuations much more accurately.

Also to be noted, for each of the three different small prediction samples chosen to display, although on average prediction performance was rather low, two of the seven goals predicted rounded to the correct value, which rounds to 28.6% prediction accuracy of the games tested.

In summary, the best linear model found through Backward Elimination is comprised of 5 predictor variables: Corsi-For, Corsi-Against, Corsi-For_Percent, Power-Play_Percent, and Losses. Relating to the class of predictors, Shooting Statistics make up the majority of the model consisting of three Corsi variables, with a sole Penalty Statistics (Power-Play_Percent) and Basic Game Statistic (Losses). From this model, the best R^2 statistic was found to be 11.61% for the in-sample data, which is

the percentage of variance for the random sample of Goals-For that is predicted by the predictor variables. A residual analysis was then performed, that found there to be a linear relationship between the predictors and a fairly constant variance across the regression line. Next a model assessment was performed that allowed a comparison of prediction estimates to actual goal scored, citing relatively low success although finding a couple of accurate estimates. Lastly, a leave-one-out cross-validation confirmed the model's low predictive power, producing poor values for the Root Mean Square Error and the Coefficient of Determination R^2.

5.2 Hockey Analytics Moving Forward

After completion of this research paper's analysis and assessment of that analysis, interesting questions regarding this paper's discovered model are brought to light. To start, in noticing that the Corsi predictors make up the majority of the fitted model, where recalling Corsi to be even shots on net for a given team *including* blocked shots, it would be interesting to further research into which defensive variables (goaltender statistics, defensive players statistics, etc.) correlate well with the goals for response. Searching out significant variables in this manner, a more explicit model built on a per-game basis with player data pertaining to each game would almost surely produce a more powerful predictive model. Also of note, as the prediction errors were rather large, it would be of interest to apply the Ridge Regression procedure mentioned in the abstract, as the use of a small biasing constant in search for better prediction estimates would be beneficial.

Hockey analytics at its current state is able to address many questions about the game. For example, in light of inadequate data revolving around a certain aspect of the game, a Match Simulation can be implemented in attempt to fill in the missing data. Or better yet, rather than fill in

missing data with simulated estimates, the project of creating of a plethora of extra data points for any given game is underway.

The I.T. company SPORTLOGiQ has begun testing their machine learning and optical recognition software, whereby a single camera tracks all player and puck movement to gather a far greater sum of data than previously possible. This idea may even come to light in the 2020-2021 season, as the National Hockey League is considering implementing puck and player tracking technology themselves, likely through the means of a tracking chip on the players and an electronic sensor in the puck. An exciting era of hockey analytics is still to come, and with it, the potential for many more interesting discoveries.

Statistics may never stack up to measuring the warmth of a smile, but that needn't mean it can't create some along the way.

6 References:

Draper, N. R., Smith, H. (1981). *Applied regression analysis* (3rd Edition). New York: Wiley

Schervish, M. J. (1995). *Theory of Statistics*. Berlin: Springer-Verlag.

Sports Reference, LLC (2007, December). Official NHL Data. Retrieved from: https://www.hockey-reference.com/

Fisher, E. (2015, September). Sportradar Signs Deal To Become NHL's Exclusive Third-Party Data Distributor. Retrieved from: https://www.sportsbusinessdaily.com/Daily/Issues/2015/09/29/Media/Sportsradar.aspx

Sports Reference LLC (2007, December). Advanced Hockey Statistics. Retrieved from: https://www.hockey-reference.com/about/advanced_stats.html

Hart, C. (2012). *Artist Index: Chris Hart*. Retrieved from:

https://web.archive.org/web/20130817050417/http://www.sapporocityjazz.jp/en/artist/chris_hart.php

Swartz, T. B. (2018). Hockey Analytics Abstract. Retrieved from:

http://people.stat.sfu.ca/~tim/papers/statsref.pdf

Marsh, J. H., Marshall, T. (2016, December). The Canadian Encyclopedia: National Hockey League (NHL). Retrieved from:

https://www.thecanadianencyclopedia.ca/en/article/national-hockey-league

Marsh, J. H., Marshall, T. (2019, June). The Canadian Encyclopedia: Stanley Cup. Retrieved from:

https://www.thecanadianencyclopedia.ca/en/article/stanley-cup

Kasan, S. (2008, October). Off-ice officials are a fourth team at every game. Retrieved from:

https://www.nhl.com/news/off-ice-officials-are-a-fourth-team-at-every-game/c-388400

Encyclopædia Britannica, Inc. (2019). ENCYCLOPÆDIA BRITANNICA: National Hockey League.

Retrieved from: https://www.britannica.com/topic/National-Hockey-League

Prewitt, A. (2014, October). Capitals Hire Analytics Consultant. Retrieved from:

https://www.washingtonpost.com/news/capitals-insider/wp/2014/10/11/capitals-hire-tim-barnes-a-k-a-vic-ferrari-as-analytics-consultant/

McKinley, M. B. (2006). Hockey: A people's History. Toronto: McClelland & Stewart. Retrieved

from: https://archive.org/details/hockeypeopleshis0000mcki

One Alarm Publishing, LLC (2019). Paul Ysebaert Hockey Statistics and Profile. Retrieved from:

https://www.hockeydb.com/ihdb/stats/pdisplay.php?pid=5839

Hockey Analytics (2020). The Eras of the NHL. Retrieved from:

http://hockeyanalytics.com/2009/02/the-eras-of-the-nhl/

Cuthbertson, B. (2005). The Starr Manufacturing Company: Skate Exporter to the World.

Journal of the Royal Nova Scotia Historical Society (8:60).

McCurdy, B. (2014, September). Summer of Analytics. Retrieved from:

https://edmontonjournal.com/sports/hockey/nhl/cult-of-hockey/summer-of-analytics-continues-as-alberta-analytics-conference-draws-a-crowd-to-saddledome

Duhatschek E. (2018, May). Jim Corsi. Retrieved from:

https://www.theglobeandmail.com/sports/hockey/jim-corsi-the-man-behind-the-number-thats-taken-over-analysis-of-the-nhl/article23676561/

Hardy, S., Holman, A. (2019, April). Hockey's Forgotten First Analytics Revolution. Retrieved from:

https://www.washingtonpost.com/outlook/2019/04/10/hockeys-forgotten-first-analytics-revolution-and-what-it-tells-us-about-games-future/

Sports Reference, LLC (2007, December). NHL Statistics Glossary. Retrieved from:

https://www.hockey-reference.com/about/glossary.html

Schneider, J. (1997, February). Cross Validation. Retrieved from:

https://www.cs.cmu.edu/~schneide/tut5/node42.html

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann. **2** (12): 1137–1143
Abstract retrieved from: http://ai.stanford.edu/~ronnyk/accEst.pdf

Cawley, G. C., Talbot, N. L. C. (2008, June). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Retrieved from: http://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

Gareth J., Witten D., Hastie T., Tibshirani, R.. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Gulitti, T. (2019, January). NHL plans to deploy Puck and Player Tracking technology next season. Retrieved from: https://www.nhl.com/news/nhl-plans-to-deploy-puck-and-player-tracking-technology-in-2019-2020/c-304218820

7 Appendix:

7.1 Glossary

All predictor variables are given and defined on pages 4-9.

7.2 Code

All code for this project was written in R and the following functions were used to execute the necessary procedures for this project.

read.csv() – Imports data of CSV format

lm() – Used to fit linear models

summary() – Generic function that prints resulting summary of various model fitting functions

corr() – Used to calculate the correlation between two variables

(*Generally: Calculates the correlation between two points in a given parameter space)*

sample() – randomly reorders elements of inputted dataset; creates a random sample

plot() – Generic function that plots R objects

abline() – Used in adding horizontal, vertical, or regression lines to a graph

rstandard() – Computes standardized residuals

qqnorm() – Used default method of this generic function to create a normal QQ plot of the variable

qqline() – Adds a line to the QQ plot created that passes through the probabilities' quantiles

resid() – Generic function that extracts model residuals

predict() – Generic function that makes predictions

length() – prints number of characters in a given string

as.character() – Generic function that converts objects into character values

as.double() – Generic function that coerces arguments to data type 'double'

append() – Adds elements to a given vector

trainControl() – Used method of function to control parameters for train(); namely to set up Leave-One-Out Cross-Validation (LOOCV)

(*Generally:* Generates parameters for furthering control over the creation of a model)

data.frame() – Produces data frames which, in general, are collections variables of a table or two-dimensional array like structure that share many properties of matrices and lists

train() – Used to calculate Root Mean Square Error and R^2 value of LOOCV

(*Generally:* Tunes models through the use of complexity parameters for associated optimal resampling statistic)

print() – Generic function that prints given arguments