Isolation of Latent Player Shooting Ability in the National Hockey League

James Kierans

A Thesis
in
The Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science (Mathematics) at
Concordia University
Montreal, Quebec, Canada

August 2021

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: James Kierans

Entitled: Isolation of Latent Player Shooting Ability in the National Hockey League

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair

Patrice Gaillardetz

_____ Examiner

Joshua Wyatt Smith

_____ Supervisor

Frédéric Godin

_____ Supervisor

Cody Hyndman

Approved by Graduate Program Director _____

Galia Dafni

_____ 2021 Dean of Faculty _____

Pascale Sicotte

Abstract

Isolation of Latent Player Shooting Ability in the National Hockey League

James Kierans

Most public research into player performance in the National Hockey League makes use of data published by the NHL itself, which is limited in scope. In the context of evaluating player shooting, i.e. the ability of a player to affect the probability of a shot of theirs becoming a goal, the use of public NHL data may produce a bias, as it does not include data about the movement of the puck prior to a shot. This thesis makes use of a privately tracked dataset which does include information regarding pre-shot movement, and inspects whether the use of this more refined dataset has a material impact on the analysis of player shooting talent. The probability of a shot becoming a goal given external factors (such as location, pre-shot movement, and game state) are used to estimate the probability of a shot becoming a goal, independent of the shooter's ability. Then, the impact a shooter has on the goal probability is captured as a coefficient estimate in a logistic regression model. This analysis is run over both datasets, and the results show that there is evidence that the more granular dataset eliminates pre-shot movement bias in the evaluation of player shooting ability.

Keywords: Ice Hockey, National Hockey League, Hockey Analytics, Shot Quality, Logistic Regression, K-Means Clustering, Random Forest, Empirical Bayes.

Contents

List of Figures

List of Tables

1 Introduction

Of particular interest in the realm of sports analytics is the use of numerical data to assess various facets of athletic skill. In the context of ice hockey play in the National Hockey League, we might be interested in a player's "shooting talent", that is, the ability of an individual to influence the probability of a shot attempt (defined as a player directing the puck towards the net with the intention of scoring) becoming a goal. While a domain expert such as a coach or scout may be able to subjectively evaluate a shooter's technique in delivering the puck to the net, there is benefit in delegating such task to Machine Learning (ML) algorithms, for the sake of establishing more robust and consistent player evaluations.

Among the first major developments in hockey analytics research was the definition and emphasis of metrics that quantify the amount of shots a team or player experiences in their favour or against them when on the ice. In hockey, where the object of the game is to see more goals in your favour than not in order to win the game, it is most natural to evaluate teams and players by how many more goals they see happen for them versus against them while playing in order to infer ability. For example, the "plus-minus" statistic, a metric which predates modern hockey analytics by decades, tallies the number of goals both for and against a player during their time on the ice, expressing the result as a differential [1].

The limitation of said approach and others like it, among several, is that goals are rare events in hockey; a typical NHL game features roughly five or six goals per sixty minutes of play. The sparsity of events leads to significant random variation in goal-based metrics which impedes repeatability and predictability. The proposed solution [2] to this sparsity was to shift towards shot-based metrics (such as "Corsi" [3] and "Fenwick" [4]), which tally the number of shots or shot attempts instead of goals. Shot attempts are ten times more common than goals, addressing the sparsity problem, while still correlating with goal-based on-ice results, under the principal that teams that generate many shot attempts also tend to score many goals, and teams that concede few shot attempts tend to allow fewer goals. Empirical historical data from the NHL shows that shot-based metrics exhibit greater repeatability and predictivity of future success than goal-based metrics (see Appendix A).

While the use of shot-based metrics constituted a significant breakthrough in the domain of hockey analytics research, they were limited in how much variance in NHL performance they explained [5]. A strict reliance on shot quantity metrics would assume that all shot attempts are of all equal quality, i.e. all as likely as one another to become goals. This assumption is intuitively and empirically unfounded. Not only is there significant spread in the quality of shot attempts in the NHL, there is a significant repeatable ability for NHL teams to generate shot attempts of an above or below average quality (see Appendix A). There is then value in considering the quality of shots that a team generates or allows for the purpose of explaining

and predicting results.

However, in contrast with shot quantity metrics, shot quality metrics explain a large portion of in-sample variance, but exhibit a much smaller degree of repeatability. On the basis of individual player evaluation, some research posits that there is no statistical evidence that a player can repeatably affect the quality of the shot attempts that occur during their ice time (i.e. shot attempts taken by opponents and teammates, not the player in question) [6]. The proposed conclusion from this work is not that players have no ability to affect shot quality of their teammates or opponents, but rather that the effect they have is small enough to not be easily detectable.

The study of "shot quality" at the NHL level, which is the practice assigning a metric denoting quality to a given shot, has thus received much attention in the past decade. Most of this research has considered shot quality models, ML classifiers which estimate the probability of a shot attempt resulting in a goal, using exogenous variables such as distance of the shot from the centre of the net, angle of the shot from the middle of the net, as well as other contextual factors [7], [8], [9]. Shot quality models in this thesis primarily employ random forest models. This is in contrast to the references above; Emmanuel Perry's [7] work makes use of similar predictors as those featured in this thesis, but employs logistic regression with binned shot types to account for non-linear interactions between predictors and the response logit. Work by Macdonald et al. [8] largely relies on logistic regression, but includes terms for shooter fatigue, which were found to be significant. The shot quality research by Paerels [9] again uses similar predictors to those featured in this thesis, but makes use of general additive models to model the relationship between shot distance and goal probability as nonlinear. Endogenous variables, such as a shooter's past performance, have also been explored for their ability to explain future shooting performance [10].

Primarily, these shot quality models have been derived for the purpose of evaluating how individuals and teams are able to generate or prevent scoring chances, based off their ability to control the quality of the shots that occur in a game. This was historically conceived as a way to bridge the gap between shot quantity metrics and shot quality metrics. The sum of shot probabilities taken by a team is referred to as an "Expected Goals" total (or "xG"), as it constitutes the estimated expected value of goals given the model's output, and is used as a hybrid estimation of team strength, accounting for both the quantity and quality of shots [11].

Moreover, shot quality models can be used to evaluate performances of individual players, either by summing their shot probabilities to establish their "expected" goal total, or by averaging their shot probabilities to estimate the mean probability of a given shot attempt becoming a goal [12]. The biggest obstacle in this pursuit is the amount of statistical noise involved in player evaluation, particular in the prediction of a player's shooting percentage (the ratio of goals scored to shot attempts), a metric with low repeatability in

small samples.

The ability to explain and predict player shooting percentage is then of interest because of its large contribution to variance in goal scoring results amongst players and teams, and for the sake of measuring the latent variable that is a player's shooting technique without having to gather additional experimental data (i.e. using only data from NHL games). Towards this end, work has been done by McCurdy [10] using generalized ridge regression to isolate player ability in the probability of a shot attempt becoming a goal. Other works have attempted to isolate latent shooting ability using repeatability thresholds such us the Kuder-Richardson-21 formula for the regression of shooting percentage results [13], [11], [14]. Empirical Bayesian analysis using a Weibull distribution as a prior belief was used by Galamini [15] in the estimation of latent goalscoring ability, which serves as the inspiration for the use of empirical Bayesian analysis of shooting percentage in Chapter 4. Other research [16] has explored the use of smoothed nonlinear spatial maps to evaluate goaltending performance at the NHL level.

Shot quality constitutes a frontier in hockey analytics research, in that it has potential to offer fresh insight into team and player performance, but presents a low degree of predictability that further research might be able to address. The quality of shots that a team generates or allows can be either inferred from past results, e.g. from shooting percentage or save percentage (ratio of saves to shots), or, as previously mentioned, can be estimated using shot quality models, which assign a probability of each shot attempt becoming a goal based on contextual factors. This approach introduces a potential bias, as the shot quality model is developed through use of statistical learning.

The bias introduced by shot quality models in the estimation of true shot quality ability is the main area of interest in this thesis. One of the main hypotheses of this research is that the main source of bias is context about shot attempts that is not accounted for by the data publicly available from the NHL. The NHL publishes location data for every shot attempt that happens in each game, but does not provide information on whether or not that shot attempt was preceded by a pass, and where that pass may have come from on the ice. The relevance of this information is in how an NHL goalie's ability to stop a shot attempt is affected by the movement of the puck through the attacking zone before the shot attempt comes. The more the puck moves before being delivered to the net (particularly lateral movement), the more the goaltender must shift in their net to prepare, and the less likely it is that they will be set and prepared to make the eventual save.

The importance of this missing information has been championed by Stimson [17], whose research has shown that pre-shot movement significantly affects the probability of a shot attempt becoming a goal at the NHL level. Pre-shot movement data has also shown utility in predicting player scoring performance [18], which indicates that pre-shot movement data may hold promise in the prediction of player shot quality performance. The goal of this research will then be to use pre-shot movement data to isolate the aspect of

shot quality ability which relates to the talent of a player for increasing the likelihood of a shot attempt becoming a goal once all other contextual factors have been accounted for, which constitutes a novel approach to player evaluation.

The interest in isolating this talent, which we will refer to as "shooting talent", is that while player shooting percentage is inherently noisy (see Appendix A), conventional hockey wisdom dictates that certain players, informally dubbed as "snipers", possess superior shooting talent. Mathematical isolation of this talent would further serve to eliminate or at least quantify the bias present in shot quality evaluation at a team or individual level.

The objective of this research is to estimate the variable of player performance that is shooting talent, and to determine how incremental data beyond publicly provided information might affect this task. Multiple shot quality models will be constructed in order to isolate the portion of player performance variance that is explained by exogenous variables. The output of these models will be used to isolate player shooting talent using logistic regression, assigning a regression coefficient estimate to each player, representing their ability. Standard logistic regression, rather than generalized ridge logistic regression as employed by McCurdy [10], is favoured since there is less concern of overfitting; McCurdy uses shooting talent estimates as a component in his shot quality model while this work does not, thus greatly reducing the number of variables in the training of the shot quality models below. Additionally, drawing inferences from the coefficients of a ridge regression is more difficult than for a standard regression [19] since the inclusion of a ridge penalty introduces a difficulty to quantify bias to the coefficient estimate; this complicates the utility of a shooting talent metric.

This new shooting talent metric will be compared with an empirical Bayesian estimate of player shooting percentage, which will server as a benchmark for the explanatory and predictive power of the newly developed shooting talent metric. This analysis will be carried out over both datasets, and the results compared.

The structure of this thesis is as follows: Chapter 2 details the datasets used for the analyses: the public NHL dataset and the privately tracked pre-shot movement dataset. The public NHL dataset has received the most focus in the public sphere, and analyses performed with it in this research serve as a benchmark for the analogous experiments tried on the pre-shot movement dataset, which has been less accessible to public researchers, and as such has received less attention.

Chapter 3 details the development of shot quality models which are used to assign to each shot in both datasets a probability of becoming a goal. This provides an expectation of the result for each shot, which can be used to evaluate players based on the quality of their shots.

Chapter 4 evaluates the utility in using empirical Bayesian analysis in the context of player evaluation for shooting results. A prior distribution is fit onto a population of player results, which is used to update the posterior distribution for a given player's individual results over the sample in question. This analysis is

conducted on observed shooting percentage for forwards and defensemen separately.

Chapter 5 combines the results of Chapters 3 and 4, using logistic regression analysis to isolate the effect a player has on the probability of a shot becoming a goal after accounting for the estimated quality of the shot. Player shooting talent is then represented as the corresponding player shooting talent coefficient in the logistic regression. This regression is run for both datasets, for forwards and defensemen separately, using the most successful shot quality models for both as developed in Chapter 3. The repeatability and predictivity of this new shooting talent metric is then assessed, using the results from Chapter 4 as a benchmark.

2 Data

Data for this project was retrieved from two sources, resulting in two separate datasets on which analyses were performed. First, game play-by-play data was scraped from the NHL API and from NHL HTML play-by-play sheets (e.g. [20]). For each game, the NHL provides an event-by-event summary, detailing every shot attempt, which was distilled into the following variables:

Table 1: Public NHL Data Set Predictors

Variable	**Type**
Time since last recorded event	Numeric
Distance From Last Recorded Event	Numeric
Distance From Net	Numeric
Angle From Centre of the Net	Numeric
Score State (trailing, tied or leading in goals)	Categorical
Strength State (5v5, 5v4, 4v4, etc.)	Categorical
Fast Rush	Categorical
Slow Rush	Categorical
Fast Rebound	Categorical
Slow Rebound	Categorical

Shot data was scraped using a custom Python script, with data taken from the 2016-17 regular season to the 2019-2020 regular season, totalling 542,223 shots attempts that resulted in either a block, miss, a save, or a goal. These shots were taken from all 6,084 games from that time span.

Second, pre-shot movement (or "passing") data was retrieved from Corey Sznajder, an independent hockey analyst, who has tracked the same data for the 2016-17 to 2019-2020 regular seasons. The data was taken from Sznajder's Patreon page. The following information is provided for each shot in the dataset:

Table 2: Pre-Shot Movement Predictors

Variable	**Type**
Scoring Chance	Categorical
Time since last recorded event	Numeric
Screen Shot	Categorical
Royal Road	Categorical
Shot Type	Categorical
Strength State	Categorical
The number of passes that preceded the shot	Numeric
Odd-man Rush	Categorical
Score State	Categorical
Pass to the point (pass to the blue line of the offensive zone)	Categorical
One-time Shot (a shot taken immediately upon receipt of a pass)	Categorical
Stretch Pass (pass which crosses the blue line and red line)	Categorical
Pass from behind the net	Categorical
Who shot the puck	Categorical
Who passed the puck	Categorical

It should be noted that in this context, an odd-man rush is defined as a sequence of play in which multiple attacking players skating into the offensive zone to attempt to score in counter-attack against a lesser number of defending players. A "Royal Road" pass is a pass that crosses through the centre of the offensive zone. Additionally, a scoring chance is defined as a shot taken from inside the home-plate area of the offensive zone, shown below:

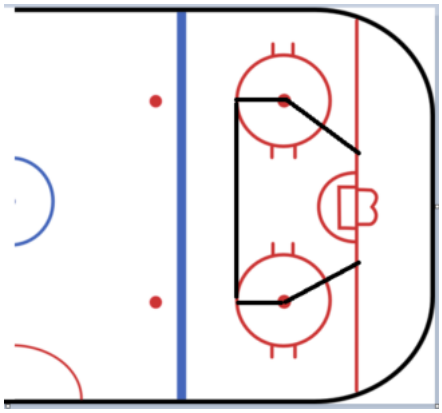

Figure 1: An illustration of the home-plate area, the area from the goal-line to the top of the face-off circles. Image courtesy of www.publicdomainfiles.com.

Rebound shots and rush shots were distinguished into two categories: fast and slow. A fast rebound shot is one that occurs within 0-2 seconds of the previous shot, and a slow rebound shot is one that occurs within 3-4 seconds. Similarly, a fast rush shot is one that occurs within 0-3 seconds of the previous event, and a slow rush shot is one that occurs within 4-5 seconds. The decision to distinguish shots in this manner, and the time intervals themselves are empirically motivated (see Appendix B), as there is a drop-off in goal probability from "fast" events to "slow" events.

Manually tracked pre-shot movement data was available for the 2016-17 to 2019-2020 regular seasons, totalling 188,017 shots, taken from 1,702 NHL games, representing 38.7% of the games played over that time span. Because this data was manually tracked by independent volunteers, getting data from every game played from the sample seasons was not possible given the workload involved.

Moreover, due to the differing origins of the two datasets, there are numerous minor inconsistencies in how events were recorded across the datasets. Most notably, time codes for each shot attempt are not recorded consistently. As a result of these inconsistencies in both recording and sample size, joining the two datasets is not feasible. Therefore, separate shot quality models have to be trained for each dataset. However, since the pre-shot movement data includes essentially all of the information provided in the public NHL data, the disadvantage arising from segregation of the datasets is minimal.

Note that for the remainder of this thesis, for simplicity, the term "shot" will be used to mean any shot

attempt that results in a block by the other team, a miss, a save, or a goal.

3 Development of Shot Quality Models

In order to isolate shooting talent at the player level, a player's shooting percentage is to be evaluated in context of the shots they take. An expectation of a player's shooting percentage given exogenous factors needs to be established, and in service of this, several shot quality models were developed using different machine learning algorithms, and using the two aforementioned datasets (NHL and pre-shot movement respectively). The goal is to use said models to explain variance in player shooting performance, under the hypothesis that the remaining variance in player shooting percentage performance will be due to shooting talent.

The intended use of the models developed in this Chapter is to serve as a basis for player evaluation. Nominally, each model is trained to assess the likelihood of a given shot becoming a goal; those estimations of shot quality can then serve to establish an "expected" shooting performance for a given player. For instance, given the quality of the shots a player took, and the number of goals they scored, we might ask whether they exceeded or failed to meet the expectations of the shot quality model that is evaluating them. For this reason, the models in this Chapter do not make use of future information relative to the player taking the shot in question: models in this Chapter may or may not incorporate information about past player performance, but future player performance is hidden from each model, so as to prevent data leakage, given that the end goal is to both explain and predict player shooting performance.

This Chapter is divided as follows: Chapter 3.1 gives mathematical background and technical details for the statistical learning techniques employed throughout the Chapter. Chapter 3.2 details the statistical learning that was performed on the NHL public dataset described in Chapter 2. Similarly, Chapter 3.3 details the statistical learning that was performed on the pre-shot movement dataset described in Chapter 2. Chapter 3.4 explores the clustering analyses of players by position and play-style, and how this affects shot quality model results. Chapter 3.5 summarizes the results of the Chapter, and provides analysis of said results.

3.1 Mathematical Background for Statistical Learning Methods

3.1.1 Methodology Employed

The following four model types were considered for learning on the public NHL data set: logistic regression, random forest, gradient boosted trees, and logistic regression/random forest ensemble. The logistic regression model was selected to accommodate the continuous and numeric nature of some of the predictors, while the tree-based methods account for the discontinuous and nonlinear nature of other predictors, as well as possible interactions. The ensemble method constitutes a hybrid approach, accounting for the smoothness of some

variables as well as the discontinuous nature of others - uniting the strengths of each type of predictor, while dampening the weaknesses of both [21]. The sub-models of an ensemble make individual predictions using simpler predictors, and aggregate their predictions in order to rise the complexity of the original task.

The decision to employ random forest as well as gradient boosted trees was justified by the fact that gradient boosted trees are more sensitive learners, but can be more prone to overfitting as compared to a random forest. There is then utility in testing both.

For the pre-shot movement data set, only the random forest and gradient boosted trees were tried, since the data set has no predictors that are both numeric and would be expected to exhibit a linear relationship with the response.

For the public NHL data set, learning was performed on 400,000 randomly selected shots (selected by random seed to ensure consistency between models). The remaining 142,223 shots were reserved as out-of-sample validation for the purpose of model comparisons. Similarly, for the pre-shot movement data set, learning was performed on 140,000 randomly selected shots, with the remaining 46,166 shots reserved as out-of-sample validation. For each model, the response is whether or not the shot in question resulted in a goal, coded as a "1" or "0".

Area under the receiver operating characteristic curve ("ROC" curve) serves as the evaluation metric for each model [22]. The ROC curve is a graphical representation of the relationship between true-positives and false-positives as the positive classification threshold is modified (in the context of classification models which output a probability and not just a class label). The area under this curve represents the overall quality of a classifier, regardless of how imbalanced the data set is towards successes (goals) or failures (saves, misses, blocks, which are far more common than goals). This is an ideal property in this context, since we aren't interested in labelling shots discretely, but rather in quantifying their quality by assigning them a probability - the purpose of a shot quality metric is more explanatory than predictive. In this sense, area under the ROC curve, denoted "AUC", is a better choice than a metric which is sensitive to imbalance in the data set, such as accuracy. An AUC score of 0.5 means the model has no discriminatory power, whereas an AUC score of 1 means the model is perfect.

For both datasets and for all models, forward variable selection was employed to determine the predictor space, i.e. a sparse model was trained first using only the most relevant predictors, and then additional predictors are added one at a time only if they provide a significant increase in model performance (as measured by "AUC"). These "relevant" predictors were the two or three with strongest variable importance scores, and that were most intuitively meaningful to the quality of a shot by conventional wisdom (a subjective judgement). This process stops when the marginal benefit of adding new predictors is negligible (defined as an increase of < 0.01 AUC). Hyperparameter tuning for each model was performed via grid search; a range

of values was determined for each hyperparameter, and then the model was trained on each combination of hyperparameters, selecting the combination which achieves the best validation score (using two-fold cross validation on the training set).

3.1.2 Logistic Regression

A logistic regression model is a generalized linear model with a logistic link function, designed primarily to handle binary classification tasks [21].

Consider first a simple linear model, with predictors $\{x_1, x_2, x_3....\}$, with target variable y, and coefficients $\{\beta_0, \beta_1, \beta_2,...\}$. Then predictions for the target variable given predictor are linear combinations of the predictors:

$$\hat{y_i} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} = \mathbf{x_i}^\top \beta \tag{3.1.1}$$

where $\hat{y_i}$ is the predicted value of a given observation y_i, and p is the dimensionality of the predictor space. Notable features of this model is that it assumes there is no significant interaction between predictors, and that predictors all have a linear relationship with the target variable. The following assumptions are not required, but are sometimes assumed to allow for seamless inference: that prediction errors exhibit constant variance, are independent of one another, and are normally distributed.

A generalized linear model is a modification of the simple linear model where a link function g is added, in order to permit for non-linear, monotonic relationships between the predictors and the target:

$$g(\hat{y_i}) = \mathbf{x_i}^\top \beta \tag{3.1.2}$$

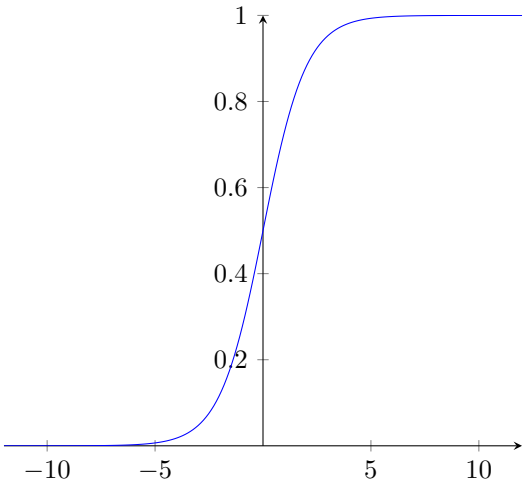

Figure 2: Logistic link function featured in logistic regression.

where g is a monotonic and differentiable function. For logistic regression, we define $g^{-1}(x) = \frac{1}{1+e^{-x}}$ (see Figure 2). Then, for binary classification, with y_i's coded as 1 or 0, we have the following formula:

$$P[y_i = 1|\mathbf{x_i}] = \frac{1}{1 + e^{\mathbf{x_i}^\top \beta}} \tag{3.1.3}$$

The logistic regression model then maps a linear combination of the predictors to a probability of a given observation being in one class or another (in this research context, the probability of a shot becoming a goal or not). The exact β which maximize the likelihood of the regression formula are found numerically, for instance using the Newton-Raphson method [23]. This calculation of likelihood relies on the assumption that responses are conditionally independent from one another given predictors, and are all Bernoulli distributed.

The advantage of a logistic regression is that its inflexibility prevents it from easily overfitting training data. Additionally, coefficients of the model are simple to interpret and confidence intervals can be obtained for all predictors. However, modelling relationships between predictors and targets as monotonic and neglecting the potential for discontinuous relationships between predictors and targets may not fit every situation. Moreover, logistic regression models do not account for interactions amongst predictors unless interaction terms are added manually, which increases dimensionality significantly.

3.1.3 Random Forests and Boosted Trees

Both random forests and boosted regression trees are generalizations of binary decision trees which use the aggregation of individual decision trees to produce more robust predictions [24], [25], [26].

A basic binary decision tree is a regression/classification algorithm which segments a predictor space

into distinct regions and predicts values for the target variable based on the characteristics of said region of the predictor space. The structure of the tree, and thus of the regions, is determined by recursive binary splitting: the training observations are initially split into two regions on a given variable x_j:

$$R_1(j,s) := \{x|x_j < s\},\ R_2(j,s) := \{x|x_j \geq s\} \tag{3.1.4}$$

For instance, a binary decision tree may split shots based on distance, segregating shots depending on whether they are closer or further than a determined threshold. The choice of j and s in a regression context are those that yield the greatest reduction in the sum of square errors formula:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1(j,s)})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2(j,s)})^2 \tag{3.1.5}$$

where $\hat{y}$ is the value predicted for observations that fall into a given region.

In a classification context, the choice of j and s are those that yield that smallest value for the Gini impurity index:

$$Gini = \sum_r \pi_r \sum_{g \in G} \hat{p}_{rg}(1 - \hat{p}_{rg}) \tag{3.1.6}$$

for the new region of the predictor space, where $\hat{p}_{rg}$ represents the portion of observations in the region r that are correctly classified, and π_r is the proportion of observations sorted into region r. If a given split is perfect, i.e. all observations are classified without error, the Gini impurity index will be zero. G_r is then a measure of the total variance across all classes g.

For each new region that is created, another binary split is performed, and this process continues until a stopping criteria is met, such as when the smallest region comprises less than a given minimum number of observations.

Binary decision trees excel when dealing with complex, highly non-linear data, since, in contrast with generalized linear models, no assumption of linearity is made. However, individual trees often exhibit a lower degree of predictive power than most other prediction algorithms, and are very sensitive to changes in data, and thus are prone to overfitting.

To preserve the strengths of tree-based methods while addressing the weaknesses, trees are often aggregated into larger learning models, such as random forests and boosted trees models.

A random forest model is an aggregation of binary decision trees, where each tree is trained on the data individually, using only a subset of the predictors chosen at random for the splitting of each node. The number of predictors used on each split is a hyperparameter chosen ad hoc for the task. The purpose in limiting

the number of predictors for each node split is that it forces trees to make recursive splits that are dissimilar from those of other trees, thus reducing the covariance between the individual predictions. The random forest model then generates predictions by averaging the outputs of the trees it is comprised of. The reduced covariance between individual predictions reduces the overall variance of the random forest prediction, which curbs the aforementioned overfitting problem. Pseudo-code for a random forest algorithm is as follows:

Algorithm 1 Random Forest

Given n, the number of predictors
Given $Ntrees$, the number of trees to be grown
Given λ, $0 < \lambda \leq n$
Define $T = \{\}$
for $1, 2, 3..., Ntrees$ **do**
  Grow tree t, restricting each split to λ randomly selected predictors
  Update: $T = T \cup t$
end for
Given p_i, the output of each tree t from T
Output: $\frac{1}{Ntrees} \sum_t^T p_i$

A boosted tree model similarly involves the aggregation of multiple decision trees. Whereas a random forest constructs trees independently, a boosted tree model trains trees sequentially. An initial tree is constructed, and its performance noted. The subsequent tree is then trained on the residuals of the prior tree, thus putting additional weight on observations that were misclassified. All trees are trained in this sequential manner, and the predictions of these trees aggregated to form a final prediction. The benefit to this incremental learning is that the algorithm can adapt more flexibly to patterns in the training data. Pseudo-code for a boosted tree algorithm is as follows:

Algorithm 2 Boosted Tree Model

Given X, the predictor space
Given $Ntrees$, the number of trees to be grown
Given d, the maximum number of splits for each tree
Given y, the vector of target values
Given $\lambda > 0$
Define: $T = \{\}$
Define: $r = y$
for $1, 2, 3..., Ntrees$ **do**
  Grow tree t, fit to data (X, r)
  Define $\hat{y}$, the vector of predictions generated by tree t
  Update: $T = T \cup t$
  Update: $r = r - \lambda \cdot \hat{y}$
end for
Given p_i, the output of each tree t from T
Output: $\frac{1}{Ntrees} \sum_t^T p_i$

It should be noted that while the use of aggregation in predictions of both random forest and boosted tree models is important to prevent overfitting, it also significantly reduces the interpretability of said models, as it is impractical to personally evaluate the individual splits of hundreds of aggregated decision trees. As such, methods for assessing variable importance are usually performed post hoc.

For regression trees, variable importance is established by finding the total reduction in RSS caused by splits performed on each given variable, where

$$RSS = \sum_{j}^{J} \sum_{i \in R_j} (\hat{y}_i - y_i)^2 \tag{3.1.7}$$

where y are the observed values, $\hat{y}$ are the predicted values, and R is one of J new regions created by the split [26]. For each variable, the total reduction in RSS across every split on that variable constitutes the variable importance score for that variable. For a classification task, a similar procedure is carried out, except the Gini impurity index is used in place of RSS. Later in this Chapter, we refer to such methods as "Gini importance" methods.

3.2 Public NHL Dataset Model

All analyses performed for this Chapter are performed on the public NHL dataset. Random forest variable importance methods [25] were used to form an initial hypothesis of which variables were of most use in this classification task, the results of which inform which variables are tried first in subsequent forward variable selection procedures. From an initial random forest regression trained with all ten predictors, the following variable importance scores were obtained:

Table 3: Random Forest Variable Importance Scores

Variable	**Score**
Time since last recorded event	913.2
Fast Rebound	776.6
Distance From Net	722.6
Score State (trailing, tied or leading in goals)	590.0
Strength State (5v5, 5v4, 4v4, etc.)	464.6
Fast Rush	391.3
Slow Rebound	353.1
Distance From Last Recorded Event	228.7
Angle From Centre of the Net	153.6
Slow Rush	105.5

The scores above were calculated from Gini importance of each predictor. Note that the variables "Fast Rebound", "Fast Rush", "Slow Rebound", and "Slow Rush" carry the same definitions from Chapter 2.

Given their elevated importance scores and their intuitive relevance in this context, "Time since last recorded event" and "Distance From Net" were used as the initial variables for the forward variable selection procedures for subsequent models.

It should be noted that all categorical variables were dummy encoded for use in the logistic regression model, and were otherwise left to be categorical for tree-based models.

3.2.1 Logistic Regression Model

An un-penalized logistic regression model was trained, with forward variable selection employed. This model achieves an in sample AUC of 0.7557, and an out-of-sample AUC of 0.7507. The following coefficients were obtained for the following predictors, which were retained for the final model.

Table 4: Logistic Regression Model Coefficients

Variable	**Coefficient**	**t-value**	**p-value**
Intercept	-1.699	-96.17	$< 10^{-4}$
Angle	-0.004	-12.08	$< 10^{-4}$
Time Since Last Event	-0.015	-24.35	$< 10^{-4}$
Score State	0.382	82.31	$< 10^{-4}$
Distance From Net	-0.041	-78.88	$< 10^{-4}$

Informally, it can be observed that the logistic regression favours shots that are taken closer to the net, and nearest the centre of the ice; location seems to matter most.

3.2.2 Random Forest Model

A random forest model was trained, with two-fold cross-validation (for hyperparameter tuning) over 400,000 randomly selected shots, using the "ranger" package in R [27]. After tuning with a grid search using out-of-sample AUC, the number of variables to split at each node was 3, and the minimum number of samples at each leaf was 250. The following six variables were retained after a forward variable selection procedure (along with their variable importance scores using Gini importance):

Table 5: Random Forest Variables

Variable	**Importance**
Time since last recorded event	1133.2
Distance of shot from the net	1108.8
Score state	625.8
Fast rebound	573.3
Distance from last recorded event	508.5
Angle of the shot from the centre of the net	316.7
Fast rush	273.8

This model achieves an in-sample AUC score of 0.9296, and an out-of-sample AUC score of 0.8377. While this model prioritizes distance of the shot from the net much like the logistic regression model, it also favours shots where there is evidence the goaltender had less time to prepare, e.g. shots taken in quick succession from the last recorded event.

3.2.3 Boosted Tree Regression Model

A boosted tree regression model was trained, with two-fold cross-validation over 400,000 randomly selected shots, using the "bstTree" package in R [28]. After tuning with a grid search using out-of-sample AUC, the number of iterations was set at a value of 1200, the shrinkage parameter at a value of 0.005, and the max tree depth at a value of 8. The following variables were chosen using forward variable selection:

Table 6: Boosted Tree Variables

Variable	**Importance**
Distance of shot from the net	1195.1
Time since last recorded event	1084.6
Fast rebound	716.1
Score state	626.2
Angle of the shot from the centre of the net	340.3
Slow rebound	298.0

This model achieves an in-sample AUC of 0.8609, and an out-of-sample AUC score of 0.8431. The variable importance scores obtained for this set of predictors lend themselves to similar conclusions as those for the random forest model. It can also be observed that this model presents a trade-off relative to the random forest model; the random forest in-sample AUC far outperforms the in-sample AUC of the boosted tree model, while the converse is true for out-of-sample AUC (albeit to a lesser extent); the random forest offers more flexibility at the cost of overfitting.

3.2.4 Ensemble Model

One of the weaknesses of the above models is that they may struggle to properly incorporate both continuous and categorical predictors. To counter this, an ensemble model was trained, combining predictions from a random forest model and a logistic regression model.

The following continuous numeric variables were included in the logistic regression model:

Both predictors are deemed significant by the model, which is consistent with previous models.

The following variables, judged to be more categorical or non-linear in nature, were included in the random forest model (along with their importance scores after training):

Table 7: Logistic Regression Variables

Variable	**Coefficient**	**t-value**	**p-value**
Distance of shot from the net	$-4.25 \cdot 10^{-4}$	-87.59	$< 10^{-4}$
Angle of the shot from the centre of the net	$-1.64 \cdot 10^{-3}$	-24.48	$< 10^{-4}$

Table 8: Random Forest Variables

Variable	**Importance**
Time since last recorded event	907.6
Fast rebound	583.8
Score state	539.5
Slow rebound	258.7
Distance from last recorded event	234.7

That there is no overlap in model predictors is deliberate; this choice was made under the hypothesis that this would lead the models to generate dissimilar/uncorrelated predictions from one another.

The random forest model, tuned to split on 2 variables at every node, and with a minimum node size of 100 observations, achieves an in-sample AUC score of 0.8352, and an out-of-sample AUC score of 0.779.

The logistic regression model achieves an in-sample AUC score of 0.7015, and an out-of-sample AUC score of 0.6972.

In-sample predictions for both models were then fed into a second logistic regression model, trained on the actual result of the shot. The model achieves an in-sample AUC score of 0.8309, and an out-of-sample AUC score of 0.805.

The following logistic regression coefficients were found:

Table 9: Ensemble Model Coefficients

Variable	**Coefficient**	**t-value**	**p-value**
Intercept	-0.034	-61.02	$< 10^{-4}$
Logistic Prediction	0.605	65.04	$< 10^{-4}$
Random Forest Prediction	1.113	259.22	$< 10^{-4}$

3.3 Pre-Shot Movement Dataset Model

The following Chapter performs similar analyses to those found in the prior Chapter, now applied to the private pre-shot movement dataset, the goal being to compare results and assess the marginal benefit of including the more granular information for each shot.

Random forest variable importance methods [25] were used to form an initial hypothesis of which variables were of most use in this classification task. From an initial random forest regression trained with all ten predictors, the following variable importance scores were obtained:

Table 10: Random Forest Variable Importance Scores

Variable	Importance
Scoring Chance	385.68
Time since last recorded event	142.84
Screen Shot	77.85
Royal Road	71.19
Shot Type	69.98
Strength State	25.00
The number of passes that preceded the shot	20.04
Odd-man Rush	19.81
Score State	18.77
Pass to the point	14.31
One-time Shot	7.89
Stretch Pass	5.31
Pass from behind the net	4.44

The scores above were calculated from Gini importance of each predictor. Notable in the table above is that the location of a given shot is still of chief importance. However, three predictors not found in the public NHL dataset, "Screen Shot", "Royal Road Pass", and "Shot Type", do provide significant predictive power. It should also be noted that a "Royal Road Pass" is defined as a pass which crosses the centre of the offensive zone. Given their elevated importance scores and their intuitive relevance in this context, "Time since last recorded event", "Scoring Chance", and "Royal Road" were used as the initial variables for the forward variable selection procedures for subsequent models.

3.3.1 Random Forest Model

A random forest model was trained, with two-fold cross-validation over 140,000 randomly selected shots, using the "ranger" package in R [27]. After tuning with a grid search using out-of-sample AUC, the number of variables to split at each node was 3, and the minimum number of samples at each leaf was 500. The following variables were retained after a forward variable selection procedure (along with their variable importance scores using Gini importance):

Table 11: Random Forest Variables

Variable	Importance
Scoring Chance	490.81
Time since last recorded event	186.64
Screen Shot	102.93
Royal Road	83.70
Shot Type	68.24
Score state	24.02
The number of passes that preceded the shot	22.99

The model achieves an in-sample AUC score of 0.8578, and an out-of-sample AUC score of 0.8413. The most important variables, "Scoring chance" and "Time since last recorded event", are those in common with the public NHL dataset. The inclusion of predictors unique to the pre-shot movement dataset proved significant however, namely "Screen Shot", "Royal Road", and "Shot Type".

3.3.2 Boosted Tree Regression Model

A boosted tree regression model was trained, with two-fold cross-validation over 140,000 randomly selected shots, using the "bstTree" package in R [28]. After tuning with a grid search using out-of-sample AUC, the number of iterations was set at a value of 600, the shrinkage parameter at a value of 0.01, and the max tree depth at a value of 3. The following variables were chosen (alongside their Gini importance scores):

Table 12: Boosted Tree Variables

Variable	**Importance**
Scoring Chance	490.81
Time since last recorded event	186.64
Screen Shot	102.93
Royal Road	83.70
Shot Type	68.24
Score State	24.02
The number of passes that preceded the shot	22.99

This model achieves an in-sample AUC score of 0.8348, and an out-of-sample AUC score of 0.8412. Given that this model boasts the same variables as that of the random forest model, the same conclusions apply with regards to the importance of predictors, since variable importance is calculated using a random forest in both cases.

3.3.3 Sparse Random Forest Model

Given that the purpose in introducing a new dataset is to assess the marginal benefit in adding more granular predictors, a new model was trained using only those predictors that are unique to the pre-shot movement dataset, in the hopes that predictions from such a model would be even less correlated with those of the public NHL dataset.

The following predictors were considered (alongside their Gini importance scores):

Table 13: Sparse Random Forest Variable Importance Scores

Variable	**Importance**
Shot Type	272.048
Royal Road	155.565
Screen Shot	101.510
Pass to the Point	50.771
The number of passes that preceded the shot	40.062
Odd-man Rush	32.385
One-time Shot	18.669
Pass from behind the net	10.624
Stretch Pass	5.804

Given that the random forest model was most successful on the pre-shot movement dataset, a random forest was trained on the sparser predictor space presented above. After tuning with a grid search using out-of-sample AUC, the number of variables to split on was set to 5, and the minimum observations per node was set to 1000. The following variables were chosen (alongside their Gini importance scores):

Table 14: Sparse Random Forest Variable Importance Scores (Trimmed)

Variable	**Importance**
Shot Type	309.96
Royal Road	132.07
Odd-man Rush	95.21
Screen Shot	69.86
Pass to the Point	25.23

This model achieves an in-sample AUC of 0.768, and an out-of-sample AUC score of 0.7583. While this model is outperformed by models with access to location data, the non-trivial AUC scores it achieves suggest that the predictors introduced by the pre-shot movement dataset hold significant predictive power.

3.4 Clustering Shooters by Play-Style

Building off the work of Ryan Stimson [29] with regard to the identification of play-styles through unsupervised clustering methods, we theorize that a player's preference for shooting or passing the puck may impact the probability of their shot becoming a goal. To this end, k-means clustering was employed using variables from the pre-shot movement dataset, for the purpose of classifying players as either "passers" or "shooters". The obtained clusters then serve as an additional explanatory variable to predict the probability of a shot becoming a goal, given the type of player who took the shot.

The following variables were used to form k-means clusters for this task:

Table 15: Clustering Variables

Variable
Shots
Scoring chances
Total shot-assists (passes to a player that shoots the puck)
Shooting Accuracy (ratio of misses to shots on goal)
Shooting Percentage (ratio of goals to shots)

It should be noted that variables representing player totals (Shots, Scoring Chances, Shot-Assists) are measured per sixty minutes of ice-time for each player, so as to adjust for players playing varying numbers of games and varying numbers of minutes per game.

A value of $k = 5$ was chosen after inspection of clusters for looking most meaningful in context of this learning task. It should be noted that this choice of k is largely arbitrary, and other values of k were tried for the subsequent analysis, yielding results that were not significantly different. Separate clusters were formed for forwards and defensemen, with player statistics being taken from the 2015-2016 season to the 2019-2020 season, with a minimum of twenty games played for each player. This yielded the following groups for forwards:

Table 16: Passing Cluster Means Forwards

Group	**Shots**	**Scoring Chances**	**Assists**	**Shooting Percentage**	**Accuracy**	**Style**
1	9.60	4.43	14.44	0.16	0.61	Low Shooting
2	16.35	6.85	14.67	0.13	0.55	Shoot First
3	11.50	5.38	20.86	0.19	0.56	Pass First
4	8.58	3.57	8.93	0.12	0.64	Low Offense
5	15.63	6.33	28.53	0.20	0.51	Strong Passer

The "Style" column is an informal characterization of said cluster:

- "Low Shooting" denotes a player with a medium number of passes, but low shooting volume
- "Shoot First" denotes a player with medium offensive results, who prioritizes shooting
- "Pass First" denotes a player with medium offensive results, who prioritizes passing
- "Low Offense" denotes a player with low totals for passing and shooting
- "Strong Passer" denotes a player with medium shot totals, and exceptional passing totals

The following clusters were found for defensemen:

Table 17: Passing Cluster Means Defensemen

Group	Shots	Scoring Chances	Assists	Shooting Percentage	Accuracy	Style
1	12.49	1.25	8.76	0.06	0.43	Shoot First
2	9.42	1.03	11.75	0.10	0.51	Pass First
3	18.41	2.75	21.96	0.11	0.42	High Volume
4	13.91	2.06	16.49	0.11	0.47	Med-High Volume
5	7.78	0.751	6.44	0.07	0.51	Low Offense

The "Style" column is again an informal characterization of said cluster:

- "Shoot First" denotes a player with medium offensive results, who prioritizes shooting
- "Pass First" denotes a player with medium offensive results, who prioritizes passing
- "High Volume" denotes a player with exceptional totals for both passing and shooting
- "Med-High Volume" denotes a player with decent totals for both passing and shooting
- "Low Offense" denotes a player with low totals for passing and shooting

The clusters derived above were then input into the pre-shot movement random forest model as predictors. In order to prevent data leakage, player groupings from one year were used to predict shots for the next year (e.g. a player's grouping from 2015-2016 would be used to predict shots in 2016-2017). Players who had too few games in the prior season to qualify for a group were classified as "Other", representing 55.2% of the dataset. These "Other" players constitute then a sixth cluster, and are included in the training of the model.

This grouping data was then included as a predictor in a new random forest model, since the random forest model was the best performing on the pre-shot movement dataset. The following predictors were featured in the model (alongside their Gini importance scores):

Table 18: Random Forest With Clustering Variables

Variable	Importance
Scoring Chance	481.11
Time since last recorded event	192.44
Screen Shot	99.93
Royal Road	86.25
Shot Type	73.13
Score state	27.00
The number of passes that preceded the shot	22.32
The group of the player who took the shot	20.75

With hyperparameters of number of variables to split on each tree of 3, and minimum number of observations per node of 500, this model achieves an in-sample AUC score of 0.8667, and an out-of-sample AUC score of 0.83.

The model was then retrained with all "Other" observations removed, resulting in a training set with 60,000 observations, and a validation set of 23,621 observations. With hyperparameters of number of variables to split on each tree of 3, and minimum number of observations per node of 500, this model achieves an in-sample AUC score of 0.8638, and an out-of-sample AUC score of 0.8243.

The grouping data was then also included as a predictor in the sparse random forest model. The following predictors were selected for this model (alongside their Gini importance scores):

Table 19: Random Forest Variable Importance Scores

Variable	**Importance**
Shot Type	291.10
Royal Road	130.39
Odd-man Rush	95.21
Screen Shot	71.24
Pass to the Point	35.00
Group	26.04

After tuning with a grid search using out-of-sample AUC, the number of variables to split on was set to 5, and the minimum observations per node was set to 1000. This model achieves an in-sample AUC score of 0.78, and an out-of-sample AUC score of 0.7659. This constitutes a very slight improvement upon the original sparse random forest model. However, given how slight the benefit is to including grouping information on a shot-by-shot basis for both the random forest model and the sparse random forest model, we can't conclude that the clustering methods hold significant value. This is reinforced by the low variable importance score for the "Group" variable in both models.

3.5 Discussion of Results

3.5.1 Summary of Results

In this Chapter, nine different models were tried over two different data sets, all of which were evaluated using AUC scores. In this Chapter we introduce a different metric for evaluating model performance, "log loss" (sometimes more formally known as "Binary Cross Entropy"):

$$\text{Log Loss} = \frac{-1}{n}\sum_{i=1}^{n}(y_i \cdot \log \hat{y}_i + (1-y_i) \cdot \log(1-\hat{y}_i)) \tag{3.5.1}$$

While an AUC score evaluates the discriminatory power of a model, log loss measures the accuracy of a model's predictions by more harshly penalizing predictions that are further removed from the observed result. This is then useful in the context of handicapping the probability of individual shots becoming goals, as the purpose of this Chapter is not to strictly classify shots, but rather to quantify the quality of each shot. The log loss scores for each model are presented below alongside the corresponding AUC scores.

For the public NHL data set, the following results were obtained:

Table 20: Public NHL Data Set Results AUC Summary

Model	**In-Sample AUC**	**Out-of-Sample AUC**
Logistic Regression	0.7557	0.7507
Random Forest	0.9296	0.8377
Boosted Tree	0.8609	0.8431
Ensemble	0.8309	0.8050

Table 21: Public NHL Data Set Results Log Loss Summary

Model	**In-Sample Log Loss**	**Out-of-Sample Log Loss**
Logistic Regression	0.172	0.173
Random Forest	0.125	0.148
Boosted Tree	0.141	0.147
Ensemble	0.155	0.156

For the pre-shot movement data set, the following results were obtained:

Table 22: Pre-Shot Movement Data Set Results AUC Summary

Model	**In-Sample AUC**	**Out-of-Sample AUC**
Random Forest	0.8578	0.8413
Boosted Tree	0.8348	0.8412
Sparse Random Forest	0.7680	0.7583
Random Forest W/ Clustering	0.8667	0.8300
Sparse Random Forest W/ Clustering	0.7800	0.7659

Table 23: Pre-Shot Movement Data Set Results Log Loss Summary

Model	**In-Sample Log Loss**	**Out-of-Sample Log Loss**
Random Forest	0.166	0.174
Boosted Tree	0.170	0.174
Sparse Random Forest	0.187	0.190
Random Forest W/ Clustering	0.165	0.174
Sparse Random Forest W/ Clustering	0.186	0.190

The two models deemed most successful for their respective datasets were the random forest models. The distributions of predicted goal probabilities across both datasets are shown below:

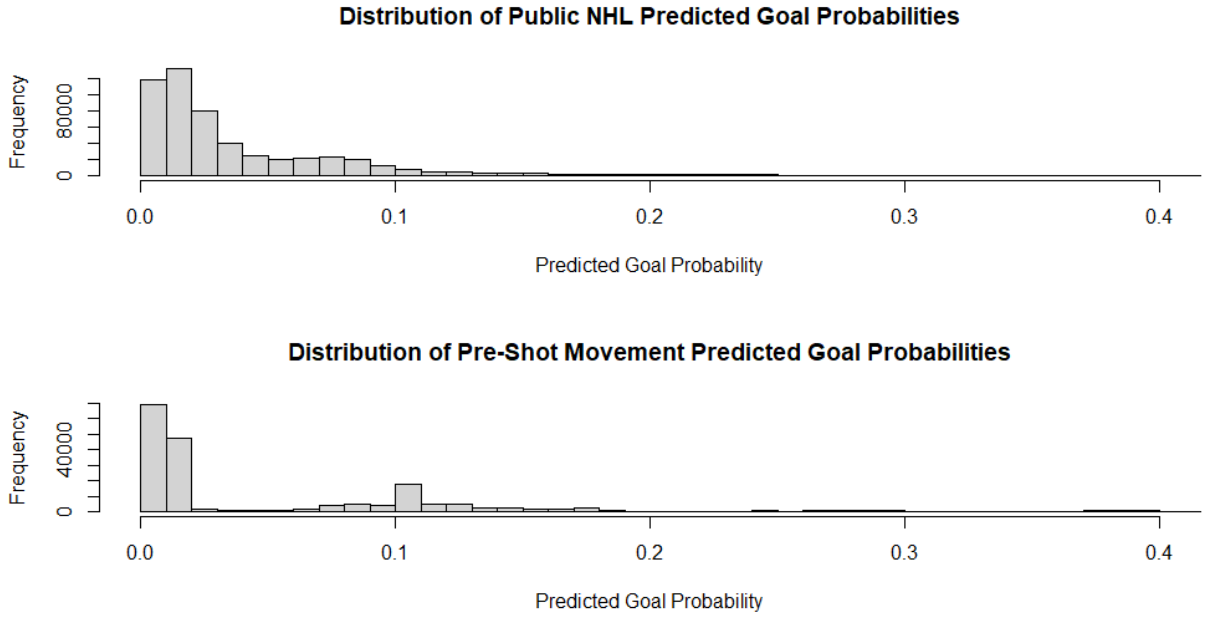

Figure 3: Histograms showing the spread of predicted goal probabilities for the random forest model trained on the Public NHL Dataset, and the random forest model (without cluster information) trained on the pre-shot Movement Dataset. Note that 0.4 does not constitute the max value of either distribution.

The pre-shot Movement predicted goal probabilities are more tightly clustered close to zero than the Public NHL dataset predicted goal probabilities. This is perhaps due to the lower sample size of the former dataset, and the lack of continuous predictors therein; given that the model has to make predictions using relatively sparse categorical variables, it would follow that the majority of predictions would be close to zero, and shots with a significant probability of becoming goals are exceptional.

For instance, one of the most significant predictors for the pre-shot Movement dataset is whether or not the goalie is being screened. However, such shots only make up 5.66% of the dataset. Similarly, only 31.22% of shots are classified as scoring chances, and 5.49% of shots are preceded by a "Royal Road" pass. Shots that don't fall into such categories are numerous, and are likely to be assigned a goal probability near zero. Those shots that are the exception and do qualify in one or more categories are given a high predicted goal value.

3.5.2 Interpretation of Results

Although the most successful models for both datasets were tree-based models and therefore difficult to interpret, we can informally note two important elements of what makes a shot more likely to become a goal.

First, the location of the shot is important. Specifically, shots that are closer to the net and nearer the

centre of the ice are more dangerous. We can infer that this might be because the shooter has a larger target to aim at, and because the goalie has a larger area to protect, and less time to react. This can be quantitatively observed through the variable importance scores for both "Distance of shot to the net", "Angle from the centre of the net", and "Scoring Chance".

Second, the less the goalie has the opportunity to prepare for the shot, the more likely it is to go in. The majority of the variables across both datasets which hold predictive power measure this directly or indirectly, e.g., the high variable importance scores for "Screen Shot", "Royal Road", and "Time since last event":

Table 24: Random Forest Variable Importance Scores Recap (Pre-Shot Movement)

Variable	**Importance**
Scoring Chance	385.68
Time since last recorded event	142.84
Screen Shot	77.85
Royal Road	71.19
Shot Type	69.98
Strength State	25.00
The number of passes that preceded the shot	20.04
Odd-man Rush	19.81
Score State	18.77
Pass to the Point	14.31
One-time Shot	7.89
Stretch Pass	5.31
Pass from behind the net	4.44

A shot being taken on the rush or being taken quickly in succession to another event indicates that a play might be developing quickly, and that the goaltender has less opportunity to set and react to the shot. A screen shot similarly indicates that a goaltender has less time to react, since they might not be able to see the release of the shot. A shot being a rebound attempt or being the result of a "Royal Road" pass (a pass which crosses the centre of the offensive zone) indicates that the goaltender has had to move quickly from one position to the next, and is thus more likely to be out of position to make the next save.

4 Assessing Player Shooting Using Bayesian Priors

One of the principal obstacles to the assessment of player shooting ability is the limited sample size that is available for analysis. The most prolific shooters in the National Hockey League take only a few hundred shots per season, and the difference between a shooting percentage that is considered adequate and one that is generally considered exceptional is minute.

The goal of this research is to evaluate a player's shooting results in the face of the uncertainty that arises from the aforementioned small sample size. While the simplest solution to a sample size issue may be to wait for a larger sample to become available, such patience may be unfeasible in practice to a decision-maker at the NHL level, who may be pressed to choose one player and their results over another in the short term. Moreover, in small sample sizes, a player's true talent level may be assumed to be constant, but over the span of years, there is evidence that a player's skill level is non-stationary [30]. Therefore, it is not necessarily suitable for the sake of accurate player evaluation to wait until a given sample size threshold is reached; decision-makers must work efficiently with the data that they do have.

In light of this problem, this Chapter makes use of Empirical Bayesian methods for parameter estimation, where the parameter in question is a player's shooting percentage. These methods allow for adjusting player talent estimates for their individual sample sizes, while also quantifying the variance in player performance by assigning a probability distribution to each player's shooting results. The key benefit to this approach is that it allows pairwise comparisons between player performances which account for the uncertainty arising from differing sample sizes - two estimated posterior distributions for a given player A and player B can be compared, and a probability calculated that one player is more skilled than the other.

The purpose of this Chapter will then be to predict future player performance, in terms of shooting percentage (ratio of goals to shots). In contrast to Chapter 3, statistical learning methods will not be employed, as future performance will be estimated using only past performance (e.g. past shooting percentage will estimate future shooting percentage). Point estimates, as well as estimated probability distributions will be derived for each player, and compared with a naive baseline in order to assess the marginal benefit of using Empirical Bayesian methods.

This Chapter is divided as follows: Chapter 4.1 provides a mathematical background for the Empirical Bayesian methods employed. Chapter 4.2 provides details of the execution of said methods for assessing player shooting percentage: the population of observed player shooting percentages is plotted, and a prior distribution fit onto the data, representing the distribution of shot probabilities on a player-by-player level in the NHL. Then, each player's distribution is updated using their observed shooting results (number of goals and number of shots), which then produces a posterior distribution for that player's shot probability.

Chapter 4.3 provides a discussion and summary of results from Chapter 4.2.

4.1 Mathematical Background for Empirical Bayesian Estimation

The purpose of Empirical Bayesian estimation is to estimate the probability distribution of a given parameter of interest, given a prior expectation for said parameter's distribution, and given experimental data presumed to be generated by said parameter [31]. A prior belief (called the prior distribution) for the distribution of the parameter is established, and that belief is updated through use of Bayes' theorem as new information becomes available:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{\sum_j P(D|H_j)P(H_j)} \tag{4.1.1}$$

where H represents our hypothesis about the distribution of the parameter, and D represents the empirical evidence about the behaviour of said parameter.

For empirical Bayesian estimate, the prior distribution is determined using empirical data. For instance, in the context of this research, the various shooting percentages of NHL forwards are plotted on a density graph, and a family of probability distributions is chosen based off the shape and behaviour of the graph. Once the type of distribution is specified, the parameters for said prior distribution are fit given the empirical data.

For example, when estimating the parameter p of a binomial distribution given empirical data, a common choice of prior distribution is a beta distribution. In this context, we can model a player's shooting percentage to be the parameter p of a binomial distribution, where the number of successes in a trial is the number of goals scored. The estimate of a player's shooting percentage if no data is available will then simply be the prior distribution, derived from the population that player is presumed to be a part of. To update the estimate of said player's performance according to new data, Bayes' theorem is used (where n is the number of shots the player took, and k is the number of goals scored):

$$P(H) = P(p = x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \text{ the prior beta distribution} \tag{4.1.2}$$

$$P(D|H) = P(k|p) = \binom{n}{k} p^k \cdot (1-p)^{n-k}, \text{ the binomial distribution} \tag{4.1.3}$$

$$B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}, \text{ the beta function} \tag{4.1.4}$$

$$P(H|D) = P(p = x|k) = \frac{P(D|H) \cdot P(H)}{\int P(D|H = y) \cdot P(H = y)dy}$$

$$= \frac{\binom{n}{k} x^{k+\alpha-1}(1-x)^{(n-k)+\beta-1}}{B(\alpha,\beta) \cdot \int_{y=0}^{1} \frac{\binom{n}{k} y^{k+\alpha-1}(1-y)^{(n-k)+\beta-1}}{B(\alpha,\beta)} dy}$$

$$= \frac{x^{k+\alpha-1}(1-x)^{(n-k)+\beta-1}}{B(\alpha+k, \beta+(n-k))}. \tag{4.1.5}$$

Therefore, the estimated distribution of the player's shooting percentage $P(H|D)$ given a prior beta distribution, and given their empirical results, is another beta distribution, with updated parameters $\alpha + k$ and $\beta + (n - k)$. We call this new distribution the posterior distribution.

The beta distribution is a common choice of distribution because it acts as a conjugate prior for a binomial distribution; if the prior distribution is beta, then the resulting posterior distribution after an update is also beta. This allows for a simple update formula: let $Beta(\alpha, \beta)$ be the prior distribution for a population of NHL players. Then for a player with n shots on goal and k goals, the posterior distribution describing their shooting percentage will be $Beta(\alpha + k, \beta + (n - k))$. Maximum likelihood fitting is used to fit the prior distribution onto the data.

Estimates of player ability in this Chapter are compared with future performance using R^2, "the coefficient of determination" of a linear regression, where:

$$\bar{y} = \frac{\sum_i y_i}{n} \tag{4.1.6}$$

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 \tag{4.1.7}$$

$$SS_{total} = \sum_i (y_i - \bar{y})^2 \tag{4.1.8}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}} \tag{4.1.9}$$

for a regression with n observations denoted with subscript i, where y_i are the response values, and $\hat{y}_i$ are the fitted regression values. R^2 then represents how correlated two variables are, on a scale from 0 to 1. Values close to zero denote weak or no correlation, while values nearer to 1 denote strong or perfect correlation.

4.2 Bayesian Analysis of Player Shooting Percentage

The public NHL dataset was used for this analysis, comprising 558,697 shots over four seasons. The goal was to use player data from the 2016-17, 2017-18, and 2018-19 seasons to estimate player shooting performance for the 2019-20 season. Empirical Bayesian prior distributions were established using data from the first three seasons, and then posterior distributions were derived using said prior.

Remember that the definition of player shooting percentage is:

$$\text{Player Shooting Percentage} = \frac{\text{Player Goals Scored}}{\text{Player Shots}} \tag{4.2.1}$$

Below we can see a density plot for player shooting percentage over the first three seasons. Players with non-zero shooting percentages below 0.3 were included (so as to eliminate outliers), leaving 685 players in the sample:

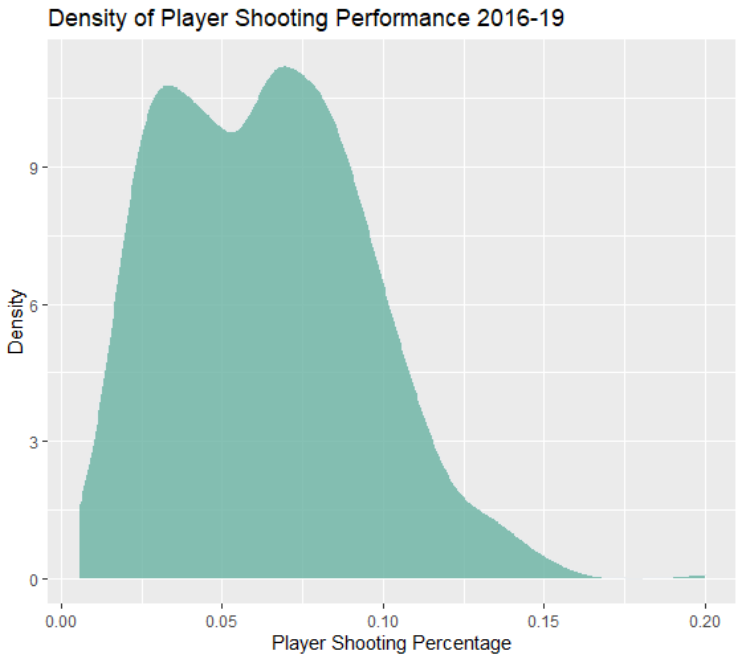

Figure 4: Density plot for player shooting percentage from the 2016-19 NHL seasons.

This empirical data exhibits bimodality, likely from the differing distributions for forwards and defensemen. Observe below a density plot for forwards only (456 players):

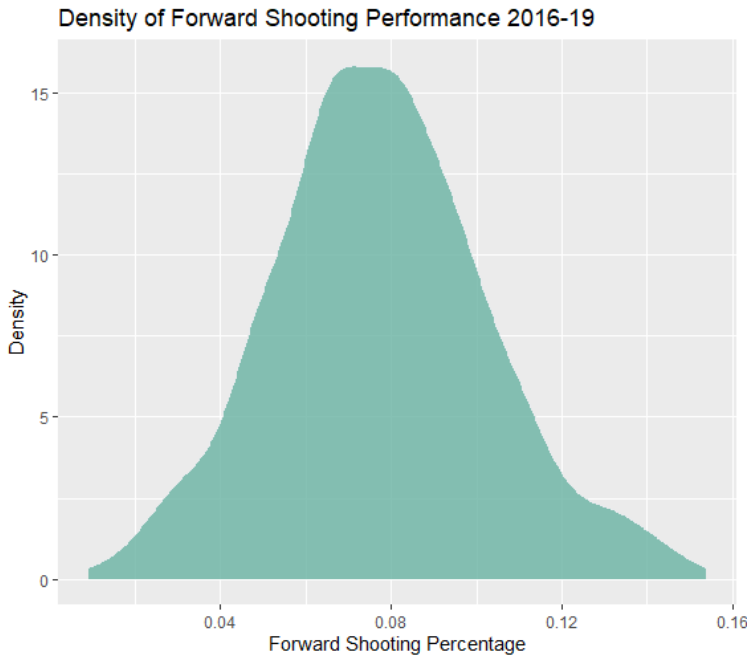

Figure 5: Density plot for forward shooting percentage from the 2016-19 NHL seasons.

Now the density plot for defensemen (230 players):

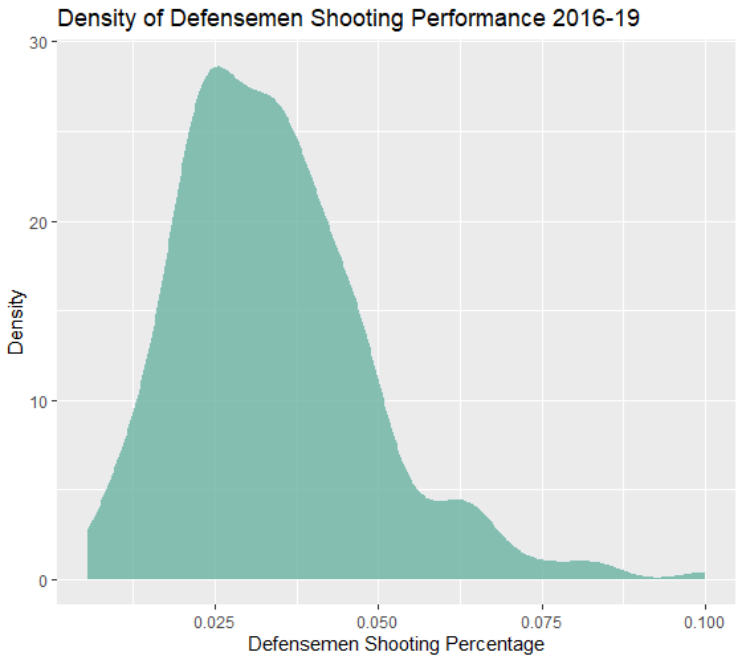

Figure 6: Density plot for defensemen shooting percentage from the 2016-19 NHL seasons.

4.2.1 Forward Shooting Percentage

A beta distribution was fit onto the forward shooting percentage data using the "MASS" R package [32] using maximum likelihood fitting, giving the following parameters: $\alpha = 12.96$ (standard error of 0.993), $\beta = 149.22$ (standard error of 11.641). For the purpose of the fit, only forwards with 200 shots or more were considered, in order to filter outliers further. This yielded the following fit for the density curve:

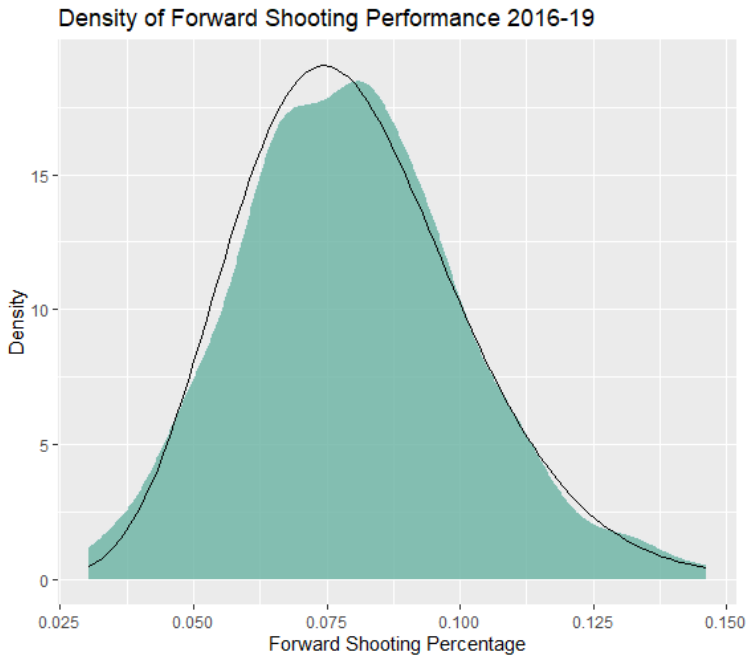

Figure 7: Density plot for forward shooting percentage from the 2016-19 NHL seasons, with a beta distribution fit over the data.

Using this prior distribution, the posterior distribution of a forwards shooting percentage is then, given k, their number of goals scored, and n, their number of shots:

$$Beta(12.96 + k, 149.22 + (n - k)) \tag{4.2.2}$$

This gives the following point estimate for future forward shooting percentage performance:

$$Estimate = \frac{12.96 + k}{12.96 + 149.22 + n} \tag{4.2.3}$$

Since the mean of a beta distribution takes the form:

$$\text{Mean} = \frac{\alpha}{\alpha + \beta} \tag{4.2.4}$$

This method of estimation was then compared with a naive baseline, i.e. a linear model with forward shooting percentage in 2016-19 as the sole predictor, and forward 2019-20 shooting percentage as the target. This model has an R^2 value of 0.098, and the following coefficients:

Table 25: Naive Model Coefficients Shooting Percentage (Public Data/Forwards

Variable	**Coefficient**	**t-value**	**p-value**
Intercept	0.053	11.15	$< 10^{-4}$
Prior Shooting Percentage	0.345	5.98	$< 10^{-4}$

By comparison, the R^2 between the posterior distribution point estimates for forward shooting percentage and 2019-20 forward shooting percentage is: 0.121. This constitutes an improvement upon the naive baseline.

4.2.2 Defensemen Shooting Percentage

A beta model was fit onto the defensemen shooting percentage data using the "MASS" R package [32] using maximum likelihood fitting, giving the following parameters: $\alpha = 5.69$, $\beta = 168.85$. For the purpose of the fit, only defensemen with 50 shots or more were considered, in order to filter outliers further. This yielded the following fit for the density curve:

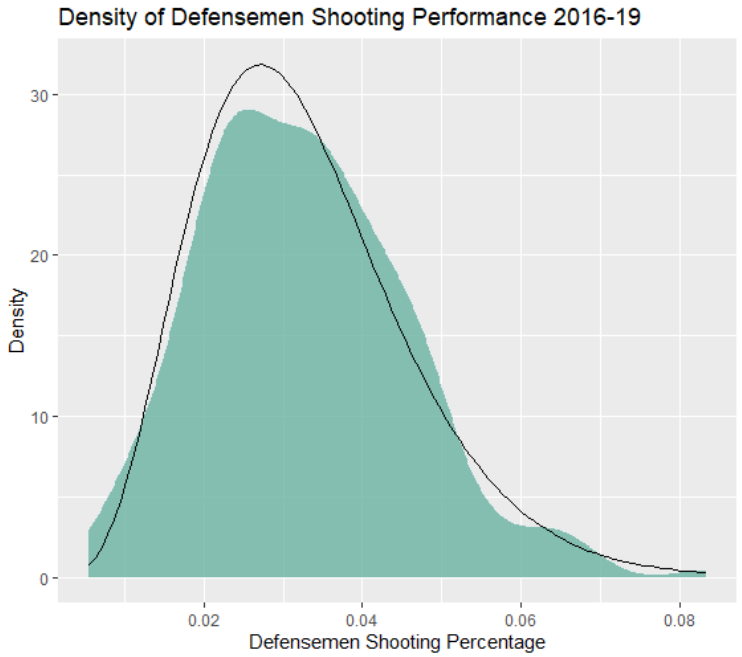

Figure 8: Density plot for defensemen shooting percentage from the 2016-19 NHL seasons, with a beta distribution fit over the data.

Using this prior distribution, the posterior distribution of a defensemen's shooting percentage is then,

given k, their number of goals scored, and n, their number of shots:

$$Beta(5.69 + k, 168.85 + (n - k)) \tag{4.2.5}$$

This gives the following point estimate for future forward shooting percentage performance:

$$Estimate = \frac{5.69 + k}{5.69 + 168.85 + n} \tag{4.2.6}$$

This method of estimation was then compared with a naive baseline, i.e. a linear model with defensemen shooting percentage in 2016-19 as the sole predictor, and defensemen 2019-20 shooting percentage as the target. This model has an R^2 value of 0.061, and the following coefficients:

Table 26: Naive Model Coefficients Shooting Percentage (Public Data/Defensemen)

Variable	**Coefficient**	**t-value**
Intercept	0.021	6.99
Prior Shooting Percentage	0.278	3.23

By comparison, the R^2 between the posterior distribution point estimates for defensemen shooting percentage and 2019-20 defensemen shooting percentage is: 0.081. This constitutes an improvement upon the naive baseline.

4.3 Summary of Results and Discussion

The purpose of this Chapter was to employ empirical Bayesian analysis for the sake of controlling for variable sample sizes for player shooting percentage evaluation. This technique improved upon the naive baseline (simple linear regression predicting next year's results) for the estimation of player shooting percentage using a beta prior. This means that the posterior distribution estimate for player shooting percentage is a good point estimator for future shooting percentage, while also providing a more robust estimation of the variance of their shooting percentage.

Such a posterior distribution can also be useful in a descriptive context. Since output of the analysis is not just a point estimate of ability but rather a probability distribution, we can robustly estimate the probability that one player has a higher true shooting percentage than another.

For example, over the 2016-19 seasons:

- William Nylander has 48 goals on 747 shots. This gives $\alpha = 60.96$ and $\beta = 848.22$.
- Brady Tkachuk has 22 goals on 275 shots. This gives $\alpha = 34.96$ and $\beta = 402.22$.

The two beta distributions representing each players' true shooting percentage can be seen below (William Nylander in blue, Brady Tkachuk in red):

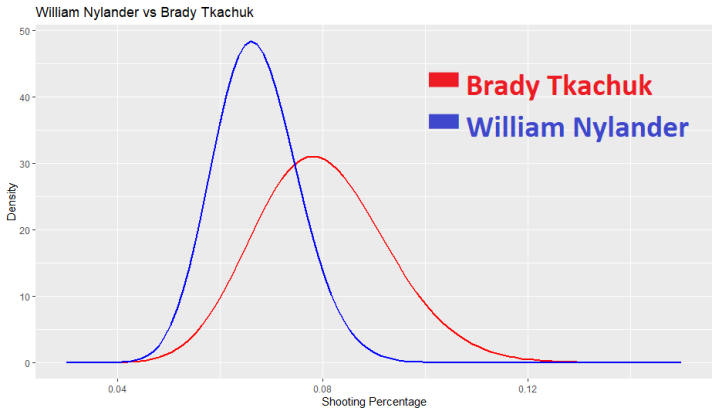

Figure 9: Density curves for William Nylander (in blue) and Brady Tkachuk (in red) shooting percentage posterior distributions.

The curve for Brady Tkachuk lies to the right of William Nylander's, but is wider, to represent the lack of certainty for his "true" shooting percentage. Using random sampling from each distribution (1,000,000 samples), there is then a 79.90% probability that Brady Tkachuk has a true shooting percentage greater than William Nylander. This style of analysis would be of use to a decision maker who needs to make pairwise comparisons between players that have uneven sample sizes.

5 Isolation of Player Shooting Talent

Isolation of player shooting talent is performed in this Chapter, using results and methods from Chapter 3 and Chapter 4. The goal is to estimate player shooting talent by measuring the impact a player has on the probability of a shot becoming a goal once the quality of the shot itself is accounted for. To this end, this Chapter employs logistic regression analysis to quantify the variance in goal probability due to "player shooting talent effects", as demonstrated by the following logistic regression formula:

$$\log \frac{p_k}{1-p_k} = \text{Intercept} + (\text{Shot Quality})_k \cdot \beta_0 + \sum_j \beta_j \cdot \mathbb{1}_{Player(k)=j} \tag{5.1}$$

where p_k is the probability associated with the k^{th} shot in the dataset, β_0 is the regression coefficient for the "Shot Quality" of the k^{th} shot, β_j is the coefficient associated with player j, and $\mathbb{1}_{Player(k)=j}$ is an indicator function returning 1 if the shot k was taken by player j, and 0 otherwise. Recall that the shot quality models selected from Chapter 3 are expressed as the probability of said shot becoming a goal given the context in which the shot is taken. This probability is agnostic of which player took the shot in question, and so shouldn't provide information to the model that overlaps with the player shooting talent coefficients.

Such a logistic regression is run on both datasets, using their respective shot quality models to assess shot quality (the random forest models from Chapters 3.2 and 3.3). Each player is assigned a coefficient β_j by the model, representing said player's shooting talent estimate. This estimate constitutes a point estimate, as well as a standard error. Under the assumptions of a generalized linear model, this estimate $\hat{\beta}_j$ converges asymptotically towards a normal distribution:

$$\hat{\beta}_j \sim N(\beta_j, \varsigma_j^2) \tag{5.2}$$

where β_j is the point estimate of the coefficient, and ς is the standard error of the j^{th} coefficient, which is calculated as:

$$\log_e L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\mathbf{X}_i'\boldsymbol{\beta}) - \sum_{i=1}^{n} \log_e(1 + \exp(\mathbf{X}_i'\boldsymbol{\beta})) \tag{5.3}$$

$$\varsigma_j^2 = \left(\frac{-\delta^2 \log_e L(\boldsymbol{\beta})}{\delta^2 \beta_j} \right)^{-1} \tag{5.4}$$

where $\log_e L(\beta)$ is the log-likelihood for the logistic regression [33], Y_i are the dependent variable values, and $\mathbf{X}$ is the design matrix.

For the purpose of establishing predictivity and repeatability of this new metric, the above regression

is performed first on the 2016-19 seasons, and then separately on the 2019-20 season. It should also be noted that much like the analyses in Chapter 4, the analyses of Chapter 5 are split between forwards and defensemen, given the bimodal nature of skater shooting performance in the NHL.

This procedure was tried on the NHL public dataset with the random forest model from Chapter 3.2, and on the pre-shot movement dataset with the random forest model from Chapter 3.3.

5.1 Public NHL Dataset

The above procedure was tried on the public NHL dataset. Since the end goal of this research is to assess the marginal benefit of using the pre-shot movement dataset, the results from this Chapter serve as a control for the results from the pre-shot movement dataset.

5.1.1 Forward Player Shooting Talent

Logistic regression was performed first to assess forward shooting talent estimates, with the regression equation proposed at the beginning of the Chapter (5.1). Only forwards with 200 shots or more were considered for the initial regression. This restriction is imposed since players with lesser shot totals were found to have coefficients that varied significantly from the population curve seen below, with standard error values much higher than their peers; this suggests this method is ill-suited to evaluate players with low sample sizes.

The regression was performed then on 288,111 shots from the 2016-19 NHL seasons, with 420 player shooting talent effect columns. Player shooting talent coefficient estimates were distributed as follows:

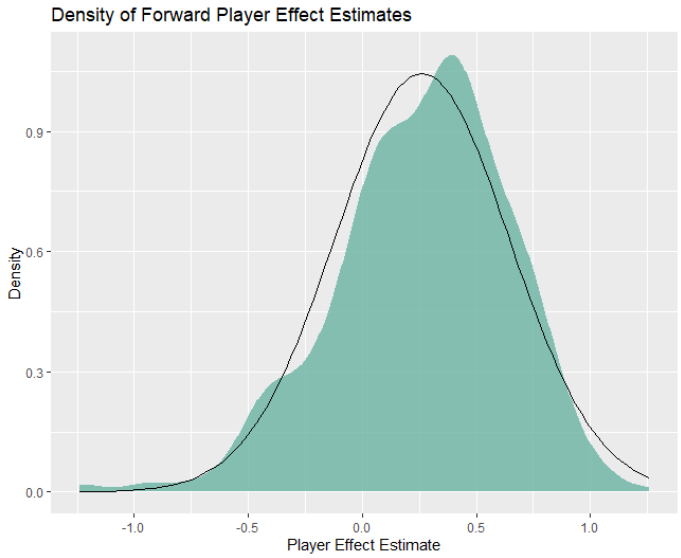

Figure 10: Density plot of player shooting talent estimates for the population of NHL forwards with 200 or more shots in the 2016-19 NHL seasons.

The player shooting talent estimates coefficients can be interpreted by how much a player can increase or decrease the "logit" in the regression equation (5.1) (recall that the "logit" is $\log \frac{p}{1-p}$). A player shooting talent estimate of 0 indicates the player has no effect on the logit, whereas a positive player shooting talent estimate indicates the player increases the probability of a shot becoming a goal (the converse is true for a negative player estimate). Notice that the distribution of player shooting talent estimates is left-skewed - the majority of players have a shooting talent estimate that is positive. This is likely due to survivorship bias; forwards with poor shooting results (either through random variance or lack of skill) tend to not last long in the NHL, and thus get filtered out before they can accrue a larger sample of shots. In practice, this means that shots in the regression that have no player shooting talent associated with them, i.e. shots taken by players not included in the sample, have a lesser chance of becoming goals.

A similar regression was run for forwards again, this time limited to the 2019-20 season (80,919 shots, 326 forwards with minimum 100 shots). Forward shooting talent coefficient estimates were compared pairwise for players with sufficient shots totals in both time frames (2016-19 and 2019-20). The R^2 between shooting talent coefficient estimates from both time frames is 0.152; this indicates that a forward having a material impact on the probability of a shot becoming a goal is a repeatable skill (as measured using this dataset). A scatter plot, with 2016-19 estimates plotted against 2019-20 estimates, illustrates this.

Moreover, the correlation between 2016-19 forward shooting talent coefficient estimates and 2019-20 observed shooting percentage was 0.0987, indicating that this measure of forward shooting talent has predictive value.

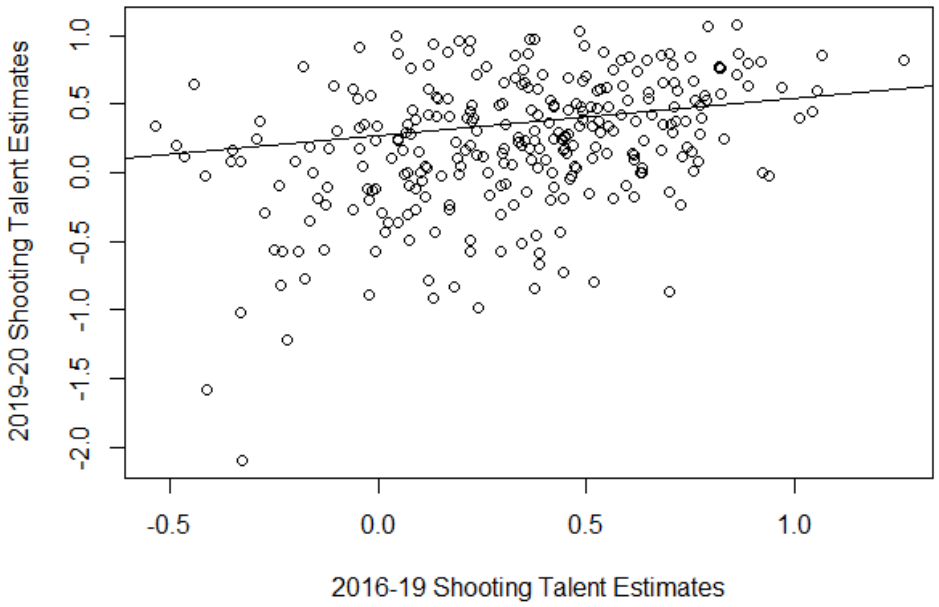

Figure 11: Scatter plot with line of best fit showing relationship between forward shooting talent estimates as measured in 2016-19 and 2019-20.

To evaluate the marginal explanatory power of shooting talent over shooting percentage results (as calculated in Chapter 4), a linear regression was performed using the following formula:

$$\text{2019-20 Shooting Percentage}_j = \text{Intercept} + \text{2016-19 Shooting Post}_j \cdot \beta + \text{2016-19 Player Shooting Talent}_j \cdot \beta_j \tag{5.1.1}$$

where the subscript j represents the j^{th} player in the sample. It should be noted that the 2016-19 "Shooting Post" values are observed shooting percentages adjusted using the beta prior distribution found in Chapter 4.2, corresponding to the mean of the posterior distribution:

$$\text{2016-19 Shooting Post}_j = \frac{12.96 + \text{Goals}_j}{12.96 + 149.22 + \text{Shots}_j} \tag{5.1.2}$$

The following regression table was obtained:

This model has an R^2 value of 0.140. As can be observed from the p-values in the table, Bayesian adjusted shooting percentage from Chapter 4 is a stronger predictor of future shooting percentage performance, and there is little benefit to the addition of shooting talent estimates to the model.

Table 27: 2019-20 Linear Regression Results (Public Data/Forwards)

Variable	Coefficient	t-value	p-value
Intercept	0.030	3.407	0.0008
2016-19 Shooting Post	0.538	3.679	0.0003
2016-19 Player Shooting Talent	0.003	0.586	0.5584

5.1.2 Defensemen Player Shooting Talent

Similar logistic regression was performed on the population of NHL defensemen in the same manner using the public NHL dataset, yielding negligible repeatability and explanatory power. The full write-up of these negative results can be seen in Appendix C.

5.2 Pre-Shot Movement Dataset

The procedure was repeated using the whole pre-shot movement dataset, with shot quality measured using the random forest model from Chapter 3.3.

5.2.1 Forward Player Shooting Talent

Logistic regression was performed first to assess forward shooting talent estimates, with the regression from Equation (5.1). The same dummy encoding is used as before, with only columns corresponding to forwards with 100 shots or more considered for the initial regression. This yielded a regression on 104,897 shots, from 241 players.

Estimates were also derived for the 2019-20 NHL season with another logistic regression, for forwards with 50 or more shots, yielding 25,810 shots from 193 players.

Forward shooting talent coefficient estimates were compared pairwise for players with sufficient shot totals in both time frames (2016-19 and 2019-20). The R^2 between shooting talent coefficient estimates from both time frames is 0.010; this indicates that a forward having a significant impact on the probability of a shot becoming a goal is not a repeatable skill (as measured using this dataset). One interpretation for this lack of repeatability is that the granular data featured in the pre-shot movement dataset allows the random forest shot quality model to capture more of the variance in shooting performance (specifically that brought about by pre-shot puck movement), leaving less variance to be captured by the player shooting talent effect estimate.

Moreover, the correlation between 2016-19 forward shooting talent coefficient estimates and 2019-20 observed shooting percentage was 0.035, indicating that this measure of forward shooting talent has little predictive value.

To evaluate the marginal explanatory power of shooting talent over shooting percentage results (as calculated in Chapter 4), a linear regression was performed using the formula of Equation (5.1.1).

The following regression table was obtained:

Table 28: 2019-20 Linear Regression Results (Private Data/Forwards)

Variable	**Coefficient**	**t-value**	**p-value**
Intercept	0.030	1.311	0.1919
2016-19 Shooting Percentage	0.550	1.996	0.0601
2016-19 Shooting Talent	0.002	0.586	0.8462

This model has an adjusted R^2 of 0.060. As can be observed from the p-values in the table above, Bayesian adjusted shooting percentage from Chapter 4 is a stronger predictor of future shooting percentage performance, and there is little benefit to the addition of shooting talent estimates to the model.

5.2.2 Defensemen Player Shooting Talent

Similar logistic regression was performed on the population of NHL defensemen in the same manner using the pre-shot movement dataset, yielding negligible repeatability and explanatory power. The full write-up of these negative results can be seen in Appendix C.

5.3 Discussion of Results

From the results of this Chapter, we can conclude that shooting talent, as measured by logistic regression coefficients capturing the effect a player has on the probability of a shot becoming a goal, does not provide marginal explanatory power beyond that which can be achieved with Bayesian adjusted shooting percentage from Chapter 4 alone.

The shooting talent metric yielded the most repeatability when measured for forwards with the public NHL dataset, and yielded negligible repeatability otherwise. It makes intuitive sense that there would be a disparity in player shooting ability even at the NHL level, but it would appear that the prediction of future player shooting results is best served by the use of Bayesian adjusted shooting percentage alone.

The curiosity driving this research was whether or not the inclusion of more granular data in a shot quality model might affect the isolation of player shooting talent in the NHL. The results from this Chapter provide evidence that the difference in datasets does affect the evaluation of shooting talent; when using the more granular dataset, shooting talent as a latent repeatable ability disappeared. This suggests that shot quality as measured using pre-shot movement data is able to account for more of the variance in shooting results, such that there is less leftover variance to be captured by the logistic regression player dummy

variables measuring player shooting talent.

This validates the original purpose of the pre-shot movement dataset, which is to explain variance in player and team performance that the public NHL dataset is not able to. This does not necessarily contradict the notion that there is a significant spread in the shooting skill of NHL players, but that such a source of variance is too minute to be detected through evaluation of goals and shots. This also suggests that the most relevant aspects of shooting skill at the NHL level might not have to do with shooting talent ("sniping" ability), but rather with the quality of shots taken by a player.

The caveat to this interpretation is that the small sample size available for analysis in the pre-shot movement dataset may be confounding the results, given that most metrics relating to the study of shot quality exhibit high variance. The high variance in player shooting performance can be observed through the varying R^2 values obtained across the two datasets. For instance, the R^2 value for predicting future shooting performance for defensemen using the public dataset (table 29) was 0.095, which is in line with the results from Chapter 4. However, under the pre-shot movement dataset, which is smaller in size, the R^2 value for predicting future shooting percentage for defensemen (table 30) was 0.1716. The disparity between these values highlights the variability in projecting shooting performance in skaters when working with smaller sample sizes.

Ideally, such an analysis could be carried out using a pre-shot movement dataset which includes all shots from all games over the span of the data collection. It's possible that the disappearance of shooting talent under the pre-shot movement dataset is the result of random variance.

Conclusion

The goal of this work was to evaluate the marginal benefit to the inclusion of a more granular dataset in the context of player shooting ability evaluation ("shooting talent"). Shot quality models were developed in order to produce an expectation for any given shot, serving the notion that the expectation could be used to evaluate a player by comparing their expected shooting results with their observed shooting results. The hypothesis was that expected shooting percentage, as calculated using public NHL data, has a bias, in that it neglects information about the movement of the puck prior to the shot. A new dataset, privately tracked and including pre-shot movement, was introduced. Shooting talent was then assessed for players in both datasets, and the results compared, to see if the inclusion of more granular data at the shot-by-shot level would assist in the isolation of shooting talent.

Chapter 3 developed the shot quality models to be used, Chapter 4 provided a benchmark against which to evaluate this proposed shooting talent metric, and Chapter 5 derived the shooting talent metric using logistic regression analysis.

Ultimately, the new shooting talent metric was not able to provide marginal explanatory power for future shooting performance beyond the benchmark set in Chapter 4. This was true over both datasets, for both forwards and defensemen. Shooting talent did exhibit repeatability for forwards as measured under the public NHL dataset, but not under the pre-shot movement dataset. This lack of repeatability is evidence of the theory that the pre-shot movement dataset does reduce bias in the evaluation of player shooting ability; that shooting talent is not repeatable in this context indicates that there is less unexplained variance in the measure of shot quality using the pre-shot movement dataset. This outcome validates the importance of pre-shot puck movement data; it should not be assumed that there is no variance arising from pre-shot movement information in the context of player shooting evaluation. However, given the amount of noise involved in projecting player performance, it cannot be ruled out that these results are due to randomness alone. Moreover, despite evidence of the significance of pre-shot movement information in the context of player shooting performance, it needs to be remarked that none of the techniques in Chapter 5 were able to improve upon the baseline set in Chapter 4 with regard to projecting future performance, which makes use of no exogenous factors.

The difficulty in explaining player shooting results then lies in the limited sample sizes available and the minute difference between an average shooting percentage result and an exceptional one. With these restrictions in mind, an avenue for further research may be to base player evaluation not on the binary outcome that is whether a shot becomes a goal or not, but whether a shot is "well taken", i.e., whether a shot is accurate, well-placed, fast, released quickly, etc.

The principle behind such an approach would be that the variance in measuring player shooting talent using goals is compounded; there is the variance in player performance to generate a "good" shot, and the variance arising from goaltender performance in stopping or allowing the shot. This latter source of variance is a particular nuisance since in theory, if the nature of the shot is controlled for, the goaltender's ability to stop it is independent of the player's ability to shoot it.

Restructuring a shooting talent analysis to focus instead on the shooter generating "good" shots rather than strictly generating goals might lead to a metric with greater repeatability and predictivity, as it would strip away variance that does not relate to the skill that is under study. The most significant obstacle to such an analysis would be the procurement of training data to perform such an analysis on. There is no publicly available database of information detailing the speed or placement of a given shot in the NHL, and to track such information, computer vision software would likely be necessary - such software is often proprietary or prohibitively expensive for public research.

References

[1]Jamie Fitzpatrick. *The Definition and Purpose of the Plus/Minus Statistic in Hockey.* URL: https://www.liveabout.com/what-is-the-plus-minus-statistic-2779372. (accessed: 04.14.2021).

[2]Rob Vollman. *Hockey Abstract Presents: ...Stat Shot, The Ultimate Guide to Hockey Analytics.* ECW Press, 2016. ISBN: 9781770413092.

[3]Wikipedia. *Corsi (statistic).* URL: https://en.wikipedia.org/wiki/Corsi_(statistic). (accessed: 04.14.2021).

[4]Wikipedia. *Fenwick (statistic).* URL: https://en.wikipedia.org/wiki/Fenwick_(statistic). (accessed: 04.14.2021).

[5]Tom Awad. *Does Shot Quality Exist.* URL: https://archive.is/Nd7BL. (accessed: 04.14.2021).

[6]Garret Hohl. *Defensemen still have no substantial and sustainable control over save percentage.* URL: https://hockey-graphs.com/2014/07/07/defensemen-still-have-no-sustainable-control-over-save-percentage/. (accessed: 04.14.2021).

[7]Emmanuel Perry. *Shot Quality and Expected Goals: Part 1.* URL: https://archive.is/5h46X. (accessed: 04.14.2021).

[8]Brian Macdonald, Craig Lennon, and Rodney Sturdivant. *Evaluating NHL Goalies, Skaters, and Teams Using Weighted Shots.* URL: https://arxiv.org/abs/1205.1746. (accessed: 05.04.2021).

[9]Taylor Paerels. "Play for the Point or Go for the Win: Expected Goals in the National Hockey League". MA thesis. California State Polytechnic University, Pomona, 2020.

[10]Micah McCurdy. *Magnus 4: xG, Shooting, and Goalie-ing.* URL: https://hockeyviz.com/txt/xg4. (accessed: 04.14.2021).

[11]Dawson Sprigings. *Expected Goals are a better predictor of future scoring than Corsi, Goals.* URL: https://hockey-graphs.com/2015/10/01/expected-goals-are-a-better-predictor-of-future-scoring-than-corsi-goals/. (accessed: 04.14.2021).

[12]Harry S. *Evaluating my Shooter xG model.* URL: http://fooledbygrittiness.blogspot.com/2018/03/evaluating-my-shooter-xg-model.html. (accessed: 04.14.2021).

[13]Harry S. *Shooter Talent and Expected Goals.* URL: http://fooledbygrittiness.blogspot.com/2018/03/shooter-talent-and-expected-goals.html. (accessed: 04.14.2021).

[14]M. W. Kuder G.F. Richardson. *The Theory of Estimation of Test Reliability. Psychmetrika.* Vol. 2. 1937, pp. 151–160. DOI: doi.org/10.1007/BF02288391.

[15]Domenic Galamini. *Comparing Scoring Talent with Empirical Bayes*. URL: https://hockey-graphs.com/2018/06/21/comparing-scoring-talent-with-empirical-bayes/. (accessed: 04.14.2021).

[16]Michael Shuckers. *DIGR: A Defense Independent Rating of NHL Goaltenders using Spatially Smoothed Save Percentage Maps*. URL: https://www.researchgate.net/profile/Michael-Schuckers/publication/267425637_DIGR_A_Defense_Independent_Rating_of_NHL_Goaltenders_using_Spatially_Smoothed_Save_Percentage_Maps/links/55251f200cf2b123c5178857/DIGR-A-Defense-Independent-Rating-of-NHL-Goaltenders-using-Spatially-Smoothed-Save-Percentage-Maps.pdf. (accessed: 05.04.2021).

[17]Ryan Stimson. *Redefining Shot Quality: One Pass at a Time*. URL: https://hockey-graphs.com/2016/01/27/redefining-shot-quality-one-pass-at-a-time/. (accessed: 04.14.2021).

[18]Ryan Stimson. *Expected Primary Points are a better predictor of future scoring than Shots, Points*. URL: https://hockey-graphs.com/2017/01/19/expected-primary-points-are-a-better-predictor-of-future-scoring-than-shots-points/. (accessed: 04.14.2021).

[19]Nimisha Chaturvedi Jelle Goeman Rosa Meijer. *L1 and L2 Penalized Regression Models*. URL: https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf. (accessed: 07.22.2021).

[20]National Hockey League. *Play By Play*. URL: http://www.nhl.com/scores/htmlreports/20202021/PL020589.HTM. (accessed: 04.14.2021).

[21]Geoffrey I. Webb Claude Sammut. *Encyclopedia of Machine Learning and Data Mining, Second Edition*. Springer, 2017. ISBN: 978-1-4899-7685-7.

[22]Tom Fawcett. *An introduction to ROC analysis. Pattern recognition letters*. Vol. 27. 8. Elsevier, 2006, pp. 861–874.

[23]Wessel N. van Wieringen. *Lecture notes on ridge regression*. 2020. URL: https://arxiv.org/pdf/1509.09169.pdf.

[24]Sreerama K. Murthy. *Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery*. Vol. 2. 1998, pp. 345–289. DOI: doi:10.1023/a:1009744630224.

[25]Leo Breiman. *Random Forests. Machine Learning*. Vol. 45. 2001, pp. 5–32. DOI: https://doi.org/10.1023/A:1010933404324.

[26]Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[27]Marvin N. Wright. *Package "ranger"*. URL: https://cran.r-project.org/web/packages/ranger/ranger.pdf. (accessed: 04.23.2021).

[28]Zhu Wang. *bst: Gradient Boosting*. URL: https://cran.r-project.org/web/packages/bst/index.html. (accessed: 04.27.2021).

[29]Ryan Stimson. *Identifying Playing Styles with Clustering*. URL: https://hockey-graphs.com/2017/04/04/identifying-player-types-with-clustering/. (accessed: 04.27.2021).

[30]CJ Turturo. *Flexible Aging in the NHL Using GAM*. URL: https://rpubs.com/cjtdevil/nhl_aging. (accessed: 06.8.2021).

[31]Jeremy Orloff and Jonathan Bloom. *18.05 Introduction to Probability and Statistics*. 2014. URL: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/.

[32]Brian Ripley, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, and David Firth. *Package 'MASS'*. URL: https://cran.r-project.org/web/packages/MASS/MASS.pdf. (accessed: 06.11.2021).

[33]Michael H. Kutner, Chris J. Nachtsheim, William Li, and John Neter. *Applied Linear Statistical Models, Fifth Edition*. McGraw-Hill Irwin, 2005. ISBN: 0-07-238688-6.

A The Repeatability and Predictivity of Common Hockey Statistics

We seek to show the repeatability and predictivity of basic shot quantity and shot quality metrics at the NHL level, illustrating the motivation for the early adoption of shot quantity metrics as tools for team evaluation, and the difficulty in quantifying shot quality.

We first define several terms:

- Corsi-For (CF): number of shot attempts (blocks, misses, saves, and goals) that a team has generated
- Corsi-Against (CA): number of shot attempts (blocks, misses, saves, and goals) that a team has allowed
- Goals-For (GF): number of goals a team has scored
- Goals-Against (GA): number of goals a team has allowed
- Shooting% (Sh%): Ratio of goals to shots on goal (saves and goals)
- Save% (Sv%): Ratio of saves to shots on goal

Of singular importance in hockey analytics research is the explanation and prediction of wins and, by extension, Goals-For and Goals-Against, which uniquely determine whether a game is won or not. In order to establish the repeatability and predictivity (w.r.t. GF and GA) of each metric, we consider every team-season from 2015-16 to 2019-20. We split the seasons into two halves by date (before or after the new year), yielding 153 team seasons split into halves.

First we examine the in-sample correlation of each metric with GF and GA (measured in R^2):

- 1st half CF vs 1st half GF: 0.3251159
- 1st half Sh% vs 1st half GF: 0.7819899
- 1st half CA vs 1st half GA: 0.3445029
- 1st half Sv% vs 1st half GA: 0.7404499

Next, the repeatability of each metric:

- 1st half CF vs 2nd half CF: 0.4396235
- 1st half GF vs 2nd half GF: 0.2351019

- 1st half Sh% vs 2nd half Sh%: 0.2351019
- 1st half CA vs 2nd half CA: 0.445053
- 1st half GA vs 2nd half GA: 0.2714957
- 1st half Sv% vs 2nd half Sv%: 0.1626525

Finally, the predictivity of each metric with respect to future goals:

- 1st half CF vs 2nd half GF: 0.139225
- 1st half Sh% vs 2nd half GF: 0.1372948
- 1st half CA vs 2nd half GA: 0.1112206
- 1st half Sv% vs 2nd half GA: 0.1719969

From these results it can be observed that shot quantity metrics (CF and CA) are more repeatable than shot quality metrics (Sh% and Sv%), and that both exhibit roughly the same degree of predictivity with goals (GF and GA). This is consistent with conventional wisdom that both are worth quantifying, but that shot quality metrics are more difficult to effectively utilize given the amount of variance they exhibit.

B Classification of Rush and Rebound Shots

In the public NHL dataset, it is possible to infer whether a shot is a "rush" attempt or a "rebound" attempt, based off of the context in which the shot was taken. A shot is qualitatively described as a "rush" shot if it is the result of a sequence of play in which multiple attacking players skating into the offensive zone attempt to score in counter-attack against a lesser number of defending players. A "rebound" shot is defined as a shot that is closely preceded by another shot.

In the context of the public NHL dataset, a "rush" shot can be inferred to have happened when a shot on net is preceded by an event in the opposite zone within a certain interval of time. A "rebound" shot can similarly be described as a shot that is preceded by another in the same zone within a given interval of time.

The choice of time intervals for both types of shots is then to be determined based off of empirical data. The following plots show the mean probability of a rush or rebound shot becoming a goal, given the specified time interval.

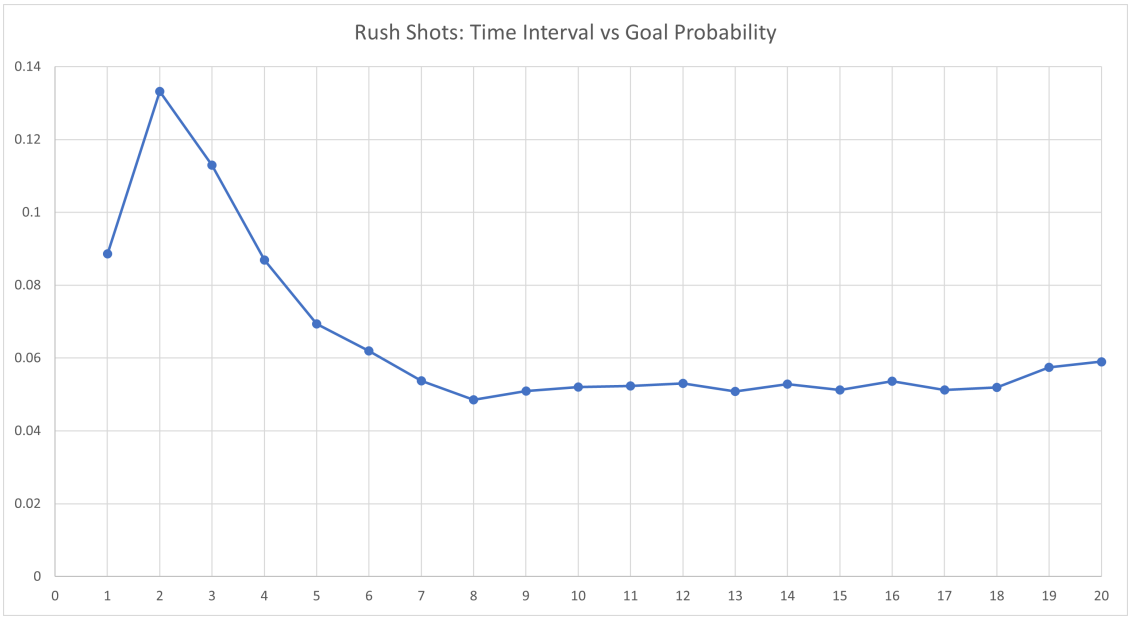

Figure 12: The frequency with which a rush shot becomes a goal given the number of seconds that has elapsed since the event that preceded it.

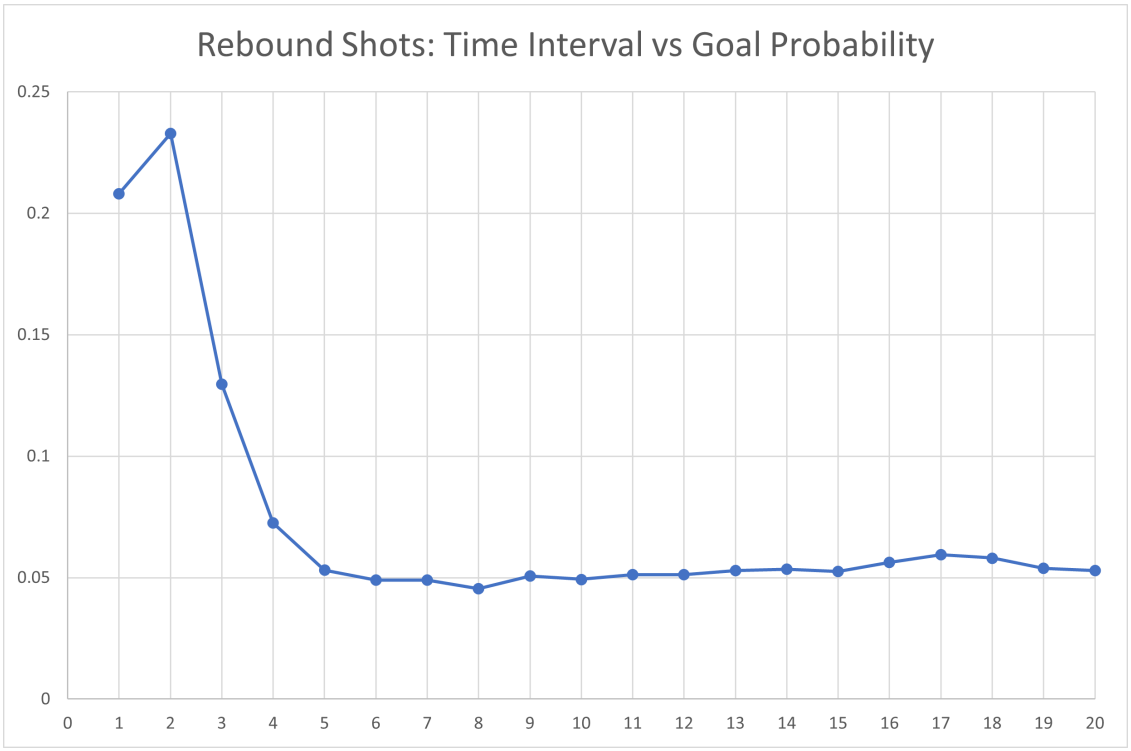

Figure 13: The frequency with which a rebound shot becomes a goal given the number of seconds that has elapsed since the shot that preceded it.

C Shooting Talent Results From Chapter 5

C.1 Public NHL Dataset

C.1.1 Defensemen Player Shooting Talent

Similar logistic regression was performed on the population of NHL defensemen. Defensemen with 200 or more shots in the 2016-19 NHL seasons were selected, yielding 147,536 shots from 225 players.

Estimates were also derived for the 2019-20 NHL season with another logistic regression, for defensemen with 100 or more shots, yielding 42,131 shots from 310 players.

Defensemen shooting talent coefficient estimates were compared pairwise for players with sufficient shot totals in both time frames (2016-19 and 2019-20). The R^2 between shooting talent coefficient estimates from both time frames is 0.0009; this indicates that a defensemen having a material impact on the probability of a shot becoming a goal is not a repeatable skill (as measured with this dataset).

Moreover, the correlation between 2016-19 defensemen shooting talent coefficient estimates and 2019-20 observed shooting percentage was 0.0001, indicating that this measure of defensemen shooting talent has little predictive value.

To evaluate the marginal explanatory power of shooting talent over shooting percentage results (as calculated in Chapter 4), a linear regression was performed using the formula from Equation (28).

The following regression table was obtained:

Table 29: 2019-20 Linear Regression Results (Public Data/Defensemen)

Variable	**Coefficient**	**t-value**	**p-value**
Intercept	0.004	0.886	0.3772
2016-19 Shooting Percentage	0.707	3.833	0.0002
2016-19 Shooting Talent	-0.001	-1.460	0.1466

This model has an R^2 value of 0.095. As can be observed from the p-values in the table, Bayesian adjusted shooting percentage from Chapter 4 is a stronger predictor of future shooting percentage performance, and there is little benefit to the addition of shooting talent estimates to the model.

C.2 Pre-Shot Movement Dataset

C.2.1 Defensemen Player Shooting Talent

Logistic regression was performed first to assess defensemen shooting talent estimates, with the regression from Equation (5.1). The same dummy encoding is used as before, with only columns corresponding to

defensemen with 100 shots or more considered for the initial regression. This yielded a regression on 41,596 shots, from 128 players.

Estimates were also derived for the 2019-20 NHL season with another logistic regression, for defensemen with 50 or more shots, yielding 14,233 shots from 123 players.

Defensemen shooting talent coefficient estimates were compared pairwise for players with sufficient shot totals in both time frames (2016-19 and 2019-20). The R^2 between shooting talent coefficient estimates from both time frames is 0.0166; this indicates that a defensemen having a material impact on the probability of a shot becoming a goal is not a repeatable skill (as measured using this dataset).

Moreover, the correlation between 2016-19 defensemen shooting talent coefficient estimates and 2019-20 observed shooting percentage was 0.0453, indicating that this measure of defensemen shooting talent has little predictive value.

To evaluate the marginal explanatory power of shooting talent over shooting percentage results (as calculated in Chapter 4), a linear regression was performed using the formula from Equation (5.1.1).

The following regression table was obtained:

Table 30: 2019-20 Linear Regression Results (Private Data/Defensemen)

Variable	**Coefficient**	**t-value**	**p-value**
Intercept	-0.023	-1.573	0.1197
2016-19 Shooting Percentage	1.607	3.450	0.0009
2016-19 Shooting Talent	-0.011	-1.690	0.0950

This model has an R^2 value of 0.1716. As can be observed from the p-values in the table, Bayesian adjusted shooting percentage from Chapter 4 is a stronger predictor of future shooting percentage performance, and there is little benefit to the addition of shooting talent estimates to the model.