National Hockey League Prospect Evaluation: Utilizing Machine Learning to Predict Amateur Prospect Career Success

Honors Thesis

Presented in Partial Fulfillment of the Requirements for the

University Honors Program

Colorado State University

By

Aaron Graff

College of Natural Sciences

Dr. Aaron Nielsen, Statistics Department

Professor Stacy Edmondson, Statistics Department

Fall 2025

Abstract

Evaluation of prospect potential at a professional level is crucial to building a championship roster in any sport. In the National Hockey League (NHL), prospect evaluation is just as critical, but even more difficult than in other sports, due to the plethora of amateur leagues globally and the youth of most prospects. This report summarizes the process of collecting North American amateur skater data between the 2007 and 2024 hockey seasons, alongside player draft position, and corresponding career expected goals in the NHL, all obtained to analyze predictive accuracy of several machine learning algorithms in predicting career expected goals. Player physical attributes, such as height and weight, were incorporated with amateur statistics from their final season before being drafted as predictor variables. Overall, bagged trees and random forest modeling had the smallest error in expected goal predictions on the test dataset. Such a model could be further developed and used as an evaluation tool for predicting prospects' NHL career success.

Introduction

In any sport, the intuitive goal of building a successful team is acquiring talented, productive players who not only contribute during a game, but work well with teammates and continue to improve through coaching. In North American professional sports, such as the National Football League (NFL), National Basketball Association (NBA), Major League Baseball (MLB), and the National Hockey League (NHL), there are three primary methods of improving talent and team performance through roster changes: free agency, trade, and the draft. Free agency is a process beginning when players previous contracts expire, allowing all teams a chance to sign various players to their roster. Trade involves an exchange of players, draft picks, and monetary funds. The draft is a process in which prospective professional players are selected by teams, usually in a pre-determined order based on the inverse of standings (such that the worst teams get to pick first in each round, and the best team picks last, for a series of rounds). In this thesis, I will focus on the third element of improving roster quality: the draft, specifically in the NHL. In any draft, it goes without saying that player projected talent and production decreases as the draft continues; in other words, one would expect the sixth round of a draft to yield significantly fewer successful professional players than the first round. In the NHL specifically, player evaluation can be particularly difficult due to several important factors: most prospects are drafted at age eighteen (compared to NFL prospects, which are typically drafted around age twenty-two after playing at a collegiate level), hockey prospects come from around the globe (particularly North America and Europe, but the game is spreading to other nations), and prospects play in a variety of leagues, with different statistical trends and levels of competition. For this reason, professional hockey teams invest significant financial and

personnel resources into prospect scouts and analytics, hoping to gain an edge in player evaluation. This is particularly important in the later rounds of the draft, when prospects are less scouted and very talented players are harder to find. Furthermore, an infinite number of variables influence how professional careers play out for all prospects, making this projection even more challenging.

One method aimed at assisting talent evaluations of scouts and coaches is machine learning. Through machine learning, it is possible to estimate the success of a prospect at the NHL level, based on any number of physical attributes and abilities, as well as amateur level statistics. While not a perfect measurement of player ability or potential, machine learning creates a model that can aggregate an overwhelming number of player statistics and data into a simplified, albeit flawed, prediction of success. This report focuses on developing several models for player career success using machine learning, with a variety of available statistics for prospects prior to being selected in an NHL Entry Draft. The ultimate goal of creating these models was to build a tool applicable to future draft classes as an additional evaluation method available to NHL scouts and general managers.

Background and Prior Research

With the number of resources and the growing market (particularly in the United States) of the NHL, it should come as no surprise that extensive prior studies have attempted to investigate this problem of predicting player success, as well as prior draft patterns and trends. In a 2024 research paper authored by three European professors in Portugal, Spain, and Denmark, unsupervised machine learning was used with K-means clustering to interpret what variables are most important in determining player success. K-means clustering is a type of

machine learning algorithm that accounts for some number (K) of points closest to another to form mutually exclusive clusters, or groups. The unsupervised nature of this algorithm indicates that the researchers did not use predictor variables to estimate a particular response variable; instead, they investigated patterns within the predictor variables themselves through clustering. In their findings, the researchers found that several metrics of practice intensity were highly correlated, and therefore redundant, such as skate acceleration and heart rate, and concluded that player training data collection should focus on training goals themselves (Rago, Fernandes, and Mohr). While this paper does not include unsupervised machine learning, these findings were relevant by speaking to idea that physical body performance trackers during player training might not be as useful in predicting player success as performance in hockey training drills specifically. In addition to this research, there is a significant amount of literature on bias not only in the NHL draft, but also in how teams utilize players based on when they were drafted. A 2013 analysis of selection bias published by Robert Beaner and Aaron Lowen, two University of Grand Valley State professors, as well as Stephen Cobley from the University of Sydney in Australia, discovered a pervasive trend of relative age effect in terms of when players were selected in NHL drafts between 1980 and 2006. Since players are typically drafted into the NHL at age 18, the researchers discovered that, on average, players born in the third and fourth quarters of their age group (i.e., July-December) were drafted forty spots later than their amateur league statistics warranted. Additionally, they concluded that those same players were about twice as likely to reach significant career success milestones, like 200 NHL points, or 400 games played (Deaner, Lowen, and Cobley). This is a fascinating discovery, considering how well-known the concept of age-related bias is, and increased evidence of the need to include birth date (or least month) in this career success analysis. In 2021, two researchers from York University in

Toronto examined the sunk cost fallacy relative to NHL player draft position in the first and second rounds. The sunk-cost fallacy is a human bias in which individuals continue to invest resources in a particular endeavor, despite knowingly receiving diminishing value, simply because they already invested resources into it in the past. By examining five-on-five ice time per game from the 2007 to 2014 NHL seasons and expected goals for and against while each player was on the ice, these researchers were able to use linear regression to estimate that draft round explained about 3.1 percent of the variation in time on ice for each player. Considering that only two draft rounds were compared, this is a noticeable explanation in variance, albeit in a simple model, that further emphasizes the desire of teams to invest playing time on a highly drafted prospect. Due to this desire to give highly drafted players more playing time, it is even more important to draft players in order of their potential to become key team contributors. Furthermore, several machine learning models examined in this thesis were used to predict next-season NHL player injuries, in a 2020 paper composed by several researchers in academic departments of Orthopedic Surgery throughout the United States, as well as the Cleveland Clinic's department of machine learning. While medical injury risk is not related to the predictive goals of this analysis, the use of random forest and XGBoost modeling aligned with usable models for analysis, with this report noting the predictive success of XGBoost to be the best (Luu, Wright, Ramkumar, et., al). Thus, XGBoost was included as a machine learning algorithm to predict player success in this research. Finally, two research papers that related most directly to the player success prediction goals of this paper involved evaluating accuracy of draft position in predicting career potential. In the first paper, researchers from the School Kinesiology and Health Science of York University, Toronto studied the relationship between draft round and games played in four major U.S. professional sports (basketball, football,

baseball, and hockey). In their findings, there was a relationship between draft round and playing time in each league, but the difference in playing time between each draft round decreased by round, particularly in the NHL (i.e., the difference in playing time between a first round pick and a second round pick was much larger than the difference between playing time for a sixth round pick and a seventh round pick), albeit while accounting for only a small percent of total variance in overall playing time (Koz, Fraser-Thomas, and Baker, 2011). This research emphasizes the difficulty of the later draft process and reinforces the potential utility of using machine learning as a tool to identify standout players in later rounds of the NHL draft. In the second paper, a 2020 paper from York University in Toronto, more than 1,500 players were analyzed for player selection round versus future performance, finding a relatively strong and significant relationship effect of draft order on player performance (Farah and Baker). However, the researchers also found that this effect was only accurate with forwards in the first two rounds of the NHL draft, and draft round was only an accurate predictor of career performance in the first round for defensemen. This analysis reinforced the inaccuracies and struggles to predict player success later in NHL drafts. This problem exists in all sports as prior scouting and information access decreases on lower-ranked prospects, but this issue appears extremely prominent in the NHL. All in all, this research reinforces the idea that machine learning might be useful as a way to differentiate between less widely known NHL prospects who likely won't contribute much at the professional level in their careers, and ones who are expected to overachieve beyond their draft projection.

Objectives

The primary goal of this machine learning analysis was to predict player success, in the form of expected goals, through a variety of player statistics and attributes that would be available prior

to draft day. As such, all models performed on this dataset were supervised (contained a response variable). Within this, a secondary goal was to identify key indicators of player success through examining variable importance, attempting to see if multiple models tend to identify similar variables that contributed to improving model predictions.

Methods

Data Collection and Preparation

For data cleaning, I utilized both Excel and R. Unfortunately, there is not a consistent database that collects all desired data for player prospects, such as pre-draft rank, physical characteristics, and amateur statistics. Thus, my dataset was created using four different sources:

1. **NHL.com Prospect Rankings** – Official pre-draft prospect rankings from the league's site itself. This data dates to the 2008 NHL entry draft, and contains a variety of player attributes, such as prospect rank, name, position, handedness, height, weight, last amateur league and team, birthday, and birth state or province.
2. **Hockey Reference Draft Data** – Contains the actual order of prospect selection in each entry draft, as well as the team each player was drafted by, position, age, career length, and that corresponding player's up-to-date statistics in the NHL since being drafted.
3. **Elite Prospects Data** – Database of player statistics for a variety of hockey leagues, both amateur and professional, around the globe. This database was extremely useful in compiling player statistics from before they were drafted into the NHL but required a subscription to download.
4. **MoneyPuck** – Site that reports hockey analytics, models, and data from each NHL game and season. This site contains a plethora of player statistics by season, beginning with

the 2008-2009 NHL season. One of these statistics was the chosen response variable, flurry adjusted expected goals per season.

The NHL official prospect rankings were easy to load in R through the NHL.com API as a JSON object. From there, each year of prospect rankings were combined into one data frame, spanning from 2008 NHL draft prospects to 2024 prospects. Primarily, the initial cleaning for this dataset relied on a few simple functions, such as excluding goalies (since attempting to project goalie value using an expected goals scored statistic would not be very beneficial), binding prospect first and last names into one variable, and creating a factor variable indicating draft year in each year's file before combining. Another important function of cleaning this dataset dealt with player rankings. There were two possible rankings for each player: midterm and final. However, some players had missing values for one of these rankings (e.g., received a final rank, but not a midterm one). To combat this issue, rather than excluding all players that might have had a missing rank, an overall (mean) rank variable was created, in which both midterm and final rank were averaged (with missing values ignored). Finally, variables that had been mutated (such as the previously separated first and last name variables and the now-combined midterm and final rank), and redundant variables, like birth city and birth state or province, were removed. Similarly, the NHL Entry Draft data was relatively easy to compile, as it also involved downloading CSV (comma-separated files) locally and reading these into R. Cleaning these files also involved removing non-skaters from the data, as well as the associated goalie statistics columns (such as wins, losses, saves, and goals against average), and once again creating a factor variable to indicate which draft year each player was selected in. Finally, to remove issues with missing values, I replaced all "NA" player statistic entries with 0.

Preparing the EliteProspects amateur data was significantly more difficult. Without the ability to access their API, manual scraping of season-by-season data was required, which was very inefficient. Therefore, rather than trying to manually download amateur skater data for as many leagues as possible, it made logical sense to narrow down prospects analyzed to the top five most common leagues that prospects had played in the previous season. These five leagues turned out to be:

1. **WHL (Western Hockey League)** – featuring teams from primarily Western Canada, as well as parts of the Pacific Northwest United States.
2. **OHL (Ontario Hockey League)** – featuring teams from Ontario, Canada.
3. **QMJHL (Quebec Maritimes Junior Hockey League)** – featuring teams from Quebec, Canada, as well as other parts of eastern Canada.
4. **ECHL (East Coast Hockey League)** – United States amateur hockey league originating on the east coast, which has expanded to include clubs across the country.
5. **USNDTP (United States National Development Team Program)** – a league of two teams: under 17 and under 18 players, organized by USA Hockey (which is the United States Olympic hockey team governing body). This team competes against other teams of similarly aged players in the United States Hockey League (USHL) and aims to centralize training for elite youth hockey players, according to Wikipedia.

After manually downloading player statistics for each of these leagues between 2008 and 2024, all data was combined in Excel, where a season and league variable were created to differentiate data. For simplicity, only the statistics for each player in their final season prior to being drafted into the NHL were chosen, meaning Elite Prospect data ranged from the 2007-2008 hockey season to the 2023-2024 season. One issue with the raw Elite Prospects database was

that, for players who played on different teams during a single season, each player had multiple rows of data: one containing total statistics, and each other row containing just statistics for each individual team they were a member of. To maintain a single row of data per player with all their overall season statistics, it was easiest to only keep rows in which team was equal to total. Additionally, player name and position were contained in one column in the following format: "playername (position)". This consistent format allowed for an Excel formula to be used that searched for the first and last parentheses in this column, separating all the text in between into a separate variable and removing it from that column. This process created separate columns for player name and position.

Finally, combining the MoneyPuck data was mostly straightforward, as it was also contained in separate CSV files for each NHL season. Creation of aggregated flurry adjusted expected goals was also relatively simple by using a pivot table. Because player statistics by season were grouped by game situation, which is impacted by penalties (even-strength five-on-five play, shorthanded four-on-five play, or five-on-four powerplay time), selection for PlayerId, name, flurry adjusted expected goals, and game situation was necessary. PlayerId and name made up the pivot table rows, while a sum of seasonal flurry adjusted expected goals comprised the values in the columns to create career flurry adjusted expected goals. However, another data quality issue emerged through this process: players with names containing accents, umlauts, or any other kind of non-English character, were being interpreted as an unidentifiable sequence of characters in Excel. This would have created issues when creating the final dataset, because players that should have been included might not have been matched to their prospect rank or draft position. To mitigate this problem, the Excel file was cleaned by identifying special, non-alphabetic characters in the data, and performing a find and replace on all instances of said characters with

the roughly translated standard English letter, after searching for a player's correct name spelling online. While certainly not a comprehensive cleaning process (by its manual nature), most notable accented letters were identified and fixed within the Excel file and updated on the pivot table. Finally, there were some players within the Money Puck dataset who were identified differently in certain years (i.e., John Doe versus Johnny Doe). To fix this issue, a similar pivot table was used to analyze players that were duplicated for each season. When examining each player with duplicates, it was possible to manually edit cells so that said player had a consistent name throughout each season of data, eliminating the possibility of that player being incorrectly categorized as multiple when the data was aggregated by career. Thus, when the corresponding CSV file was loaded into R, most of the possible data cleaning was already completed.

Data Joining

At this point, four cleaned, but distinct, datasets existed in code: prospect rankings, NHL draft history, amateur statistics from five North American leagues, and career flurry adjusted expected goals. To create a comprehensive dataset to train and test machine learning models on, this data needed to be joined together into one table. Because the NHL official prospect rankings served as the base dataset for all other corresponding data collected, left joins needed to be performed on a common column shared by tables (which results in the columns in the second table being appended at the end, only for values that match the prospect rankings column joined upon). To increase join efficiency and cleanliness, the best approach found for this problem was to create an ID column on each of the three season-by-season datasets (rankings, draft results, and amateur statistics) that contained draft year, as well as player name. This created a unique identifier for each player in the dataset to be joined. The career expected goals by player did not require such a unique identifier, because it was an aggregate statistic for each player over their

entire career (no season-by-season element of this data was incorporated). After this ID value was added to each dataset (using R for the prospect rankings and NHL draft history data, while using Excel for the Elite Prospects and Money Puck data), joins were now possible on a common ID column for each of the three season level datasets, with career expected goals to be joined simply using player name. However, there was still a significant hurdle to consider before joining data. Because each of these datasets came from different sources, there were variations in player naming (i.e., Matthew Smith versus Matt Smith), misspellings, and unexpected characters that did not make every single player's ID match perfectly between datasets. This concern was further justified by initially performing a straight test join (using perfect matches). Perfect left joining retained less than 1,600 of the 3,790 players contained in the overall NHL prospect ranking dataset. While it was expected that not every player ranked as a prospect was drafted into the NHL, with approximately 195 players selected in each NHL draft since 2008, over 16 seasons, we would expect around 3,120 players to be matched correctly, not less than 2,000. To overcome this issue, this analysis utilized a *string distance* left join, rather than a pure left join, which is provided in the R package "fuzzyjoin". The main premise of this method is to specify a maximum difference that two strings can have for them to still be matched together as a valid left join, with number of characters different between the two strings impacting quantified distance between them. This maximum distance parameter is critical to tune properly, however, because specifying too large of a string distance relaxes the join requirements too much, leading to significant duplicate matches, even for players who already had a pure left join match previously. After experimenting with possible maximum string distances, trying to balance the desire to leave out as few prospects as possible while not creating "hallucinated" false player observations, 0.15 was settled on. While certainly not a foolproof, perfect number, this value

increased player matches by nearly 700, and did not appear to create noticeably incorrect player matches. The metric used for string distance, called Jaro-Winkler (JW), is a normalized metric giving more favorable ratings to strings matching from the beginning, while accounting for number of matching characters and overall string length. A JW value of 0 means strings match exactly, and a value of 1 means there is no match. So, allowing a maximum JW value of 0.15 allows for slight string variation, but not significant flexibility, balancing the desire for more data with the need to avoid falsely created mismatches. To further limit the risk of false mismatches, after performing this string distance left join on the ID column, only observations that had the minimum string distance were selected between the prospect rankings name variable and the NHL draft history player name variable. Theoretically, for players who already had a perfect match between ID columns, the minimum string distance for the perfect match would be zero, leading to the perfect match being selected (and preventing mismatches of already joined player data). After performing these functions and removing rows with missing values for "team drafted to" or "overall draft pick" (removing players who were not drafted), a combined data frame of prospect rank, physical attributes, actual draft position, and actual draft team was created, containing 2,475 players drafted between 2008 and 2024.

For the remaining data joins (prospect data and career expected goals), it was not logical to perform a string distance join, because amateur league data was only collected for five leagues, and certainly more players who did not play in these leagues (such as those in Europe) were drafted between that time. Performing string distance joining would risk incorrectly including those players in the model and biasing the model further. Thus, in this case, perfect left join made sense, as leaving out prospects who did not play in the WHL, OHL, QMJHL, USNTDP, or ECHL was crucial. After performing a pure left join between the combined

rankings and draft history data with the amateur statistics, 1,541 players were left, with one final dataset to join.

Finally, to add the flurry adjusted career expected goals to this aggregate dataset of prospect rankings, physical attributes, actual draft position, and amateur statistics, another pure left join was performed on the name variable present in both. This pure left join preserved all 1,541 players (every name was correctly matched), leaving this as the final size of the total dataset.

Variable Selection

With all data now collected, cleaned, and joined, the final step was to select variables for the model. All in all, twenty-one predictor variables were included, along with the one response variable (flurry adjusted career expected goals): **ID, Overall Draft Position, Overall** (Mean) **Rank, NHL team** (drafted by)**, Position, Age, Birth Month, Nationality, Shoots** (left or right handed), **Height** (inches), **Weight** (pounds), **Last Amateur League** (WHL, OHL, QMJHL, ECHL, or USNTDP), **Draft Year**, and the following amateur league stats from prospects' final year before selection: **Games Played, Goals, Assists, Total Points, Points Per Game, Penalty Minutes, Plus Minus Rating** (goals for minus goals allowed when player is on the ice)**,** and **Last Amateur Team**. For model training, ID was dropped (as a string of draft year and name should not add predictive capabilities to any model), but ID was saved for the test dataset, so that ID could be rejoined to player predictions for interpreting accuracy. Birth month was included, rather than raw date, to investigate if this variable might have any predictive power, based on the age-related player selection bias identified in one of the background research papers examined.

Additionally, it is important to define the response variable chosen for this report, in terms of what exactly it measures and why it was selected. Expected goals in the NHL is a statistic calculated by multiplying the probability of any given shot on goal going into the net, and adding together these probabilities for each shot. As an example, if a player takes two shots in a game: one with a 14 percent chance of scoring, and another with a 34 percent chance of scoring, his aggregate expected goals for that game would be 0.48 (0.14 + 0.34). To make this statistic *flurry-adjusted*, one must first understand the concept of a flurry. In hockey, there are often game situations in which one team has a significant number of shots on goal in a short period of time, called a flurry. In flurry situations, there are typically high probabilities of scoring with each shot. But, for the flurry to happen, the puck must not score (otherwise, play would stop and the puck would be dropped at center ice to reset play). Thus, adjusting expected goals for flurries discounts expected goal value for each successive shot in a flurry, because that shot would not have existed if one of the previous shots scored. Essentially, adjusting for in-game flurries "balances" or deflates the raw expected goals statistic. Because this statistic was sourced from Money Puck, it relies on Money Puck's calculation of flurry adjusted expected goals, which is as follows:

$$Flurry\ Adj.Exp\ Goal\ Value = Prob\ No\ Score\ in\ a\ Flurry\ Yet * Reg\ Shot\ Exp.Goal\ Val$$

According to MoneyPuck, this statistic is more balanced and predictive of actual player goal scoring, which is why it was selected as the response variable over raw expected goals in this analysis.

Machine Learning Methods and Model Selection

Overall, six models were chosen and fit to the training dataset in this analysis. Three models were linear regression based, and three were decision tree based.

Linear Based Models:

1. **LASSO Regression:** Combines the principles of linear regression (minimizing the residual sum of squares between observations and a linear fit) with a budget parameter, constraining how large the absolute value of each parameter can be. This parameter allows the model to minimize some coefficients to zero, eliminating them from the model entirely and acting as an intuitive variable selection process.
2. **Ridge Regression:** This method also maintains the principle of minimizing residual sum of squares (RSS), but rather than constrain the model with a budget for parameter size, it incorporates a shrinkage penalty, penalizing predictor coefficients for being too large. This parameter is also tunable, and while it does decrease the value of most predictor parameters β_j, unlike LASSO, no coefficients are completely minimized to zero, so every initial predictor variable is utilized, even in a small capacity.
3. **Mixed LASSO/Ridge Regression:** As it sounds, this method incorporates aspects of both LASSO and Ridge Regression, working to minimize RSS while including both a budget for the absolute value of each predictor coefficient and a shrinkage penalty.

Tree-Based Models (each with 1000 trees grown):

1. **Boosted Trees:** grows decision trees, where the model decides on a huge series of splits between observations in the data repeatedly before reaching a predicted value, while "learning" from previous error by adding a smaller version of each tree to the overall

model after each iteration (this learning aspect is called *boosting*). Ideally, the model will improve performance by decreasing overall training error each time it builds off previous trees until error is minimized.

2. **Bagged Trees/Foresting:** Follows the same idea of creating a series of decision trees but uses repeated samples of training data (called *bootstrapped* samples) with replacement, creating trees using each bootstrapped training sample and then averaging resulting predictions.
3. **Random Forest:** This method is very similar to bagged trees. The only difference is that, when creating trees on the bootstrapped samples created, only a random subset of predictors is allowed to be chosen for each tree split, rather than all predictors (as in a bagged fit). The idea behind this is to decrease correlation between trees before predictions are averaged, because if the model can use all predictors to split in every tree, there will likely be a lot of trees with initial splits that use the exact same variable (thus meaning trees are highly correlated). Random foresting attempts to mitigate this problem and create trees that vary, to improve averaged model predictions.

Ultimately, this set of models was chosen because linear models provide a more interpretable model concept, generally at the expense of some predictive capabilities. This acted as a baseline for model performance, with the idea that the tree-based models (which are more difficult to comprehend and interpret) would likely have more predictive accuracy on the test dataset. Additionally, choosing two sets of models that act on the same principle of linear regression and decision trees allows for easier comparison between model performance within each type of model.

Model Evaluation and Results

For all fitted models, data was split randomly into training and testing using an 80 percent to 20 percent split. This left 1,232 players in the training data set, compared to 309 players for the test dataset. These datasets remained the same for all models to allow for accurate comparisons.

Linear-Based Models

LASSO Regression

To find the optimal budget parameter for model performance, ten-fold cross validation was used prior to fitting a final model. Cross-validation works by splitting the data into k parts (in this case, ten), training on k-1 parts (10-1, or 9 in this case), and testing on the last fold remaining. This process is repeated k (again, 10 in this case) times, each time with one of the folds being treated as a test set (GeeksforGeeks). Based on the goal of minimizing root-mean squared error (RMSE), the optimal parameter budget constraint was 1, and this led to only 68 out of 170 possible variables being selected for use. Overall, RMSE on the test set was approximately 48.24, which can be interpreted as the square root in error of flurry adjusted expected goals for all 309 predicted players. Below is a scatter plot of predicted versus actual flurry adjusted expected goals for the test set on this model:

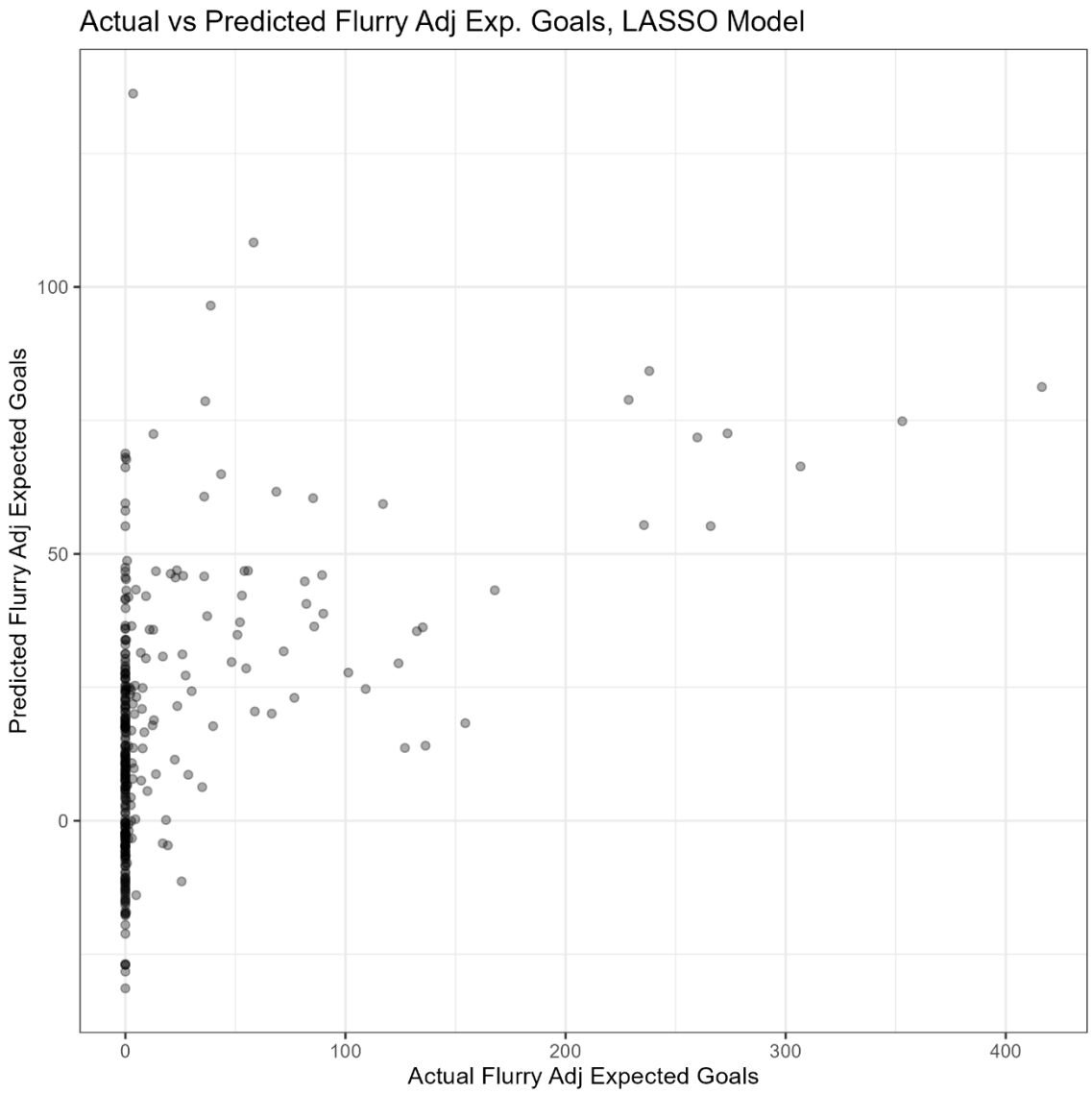

Figure 5.1 LASSO Model Actual vs Predicted Flurry Adjusted Expected Goals on the Test Set

We can see that, while there appears to be some linearity for players above zero flurry adjusted expected goals, the model consistently underestimated this statistic compared to actual values. Additionally, with such a significant set of players with no expected goals, the model struggled to make accurate predictions, with predictions for players with no actual expected goals ranging from less than -25 to above 50. In addition to examining predictions, it is possible

to make interpretations about variable importance by simply graphing variables with the largest coefficients in this model:

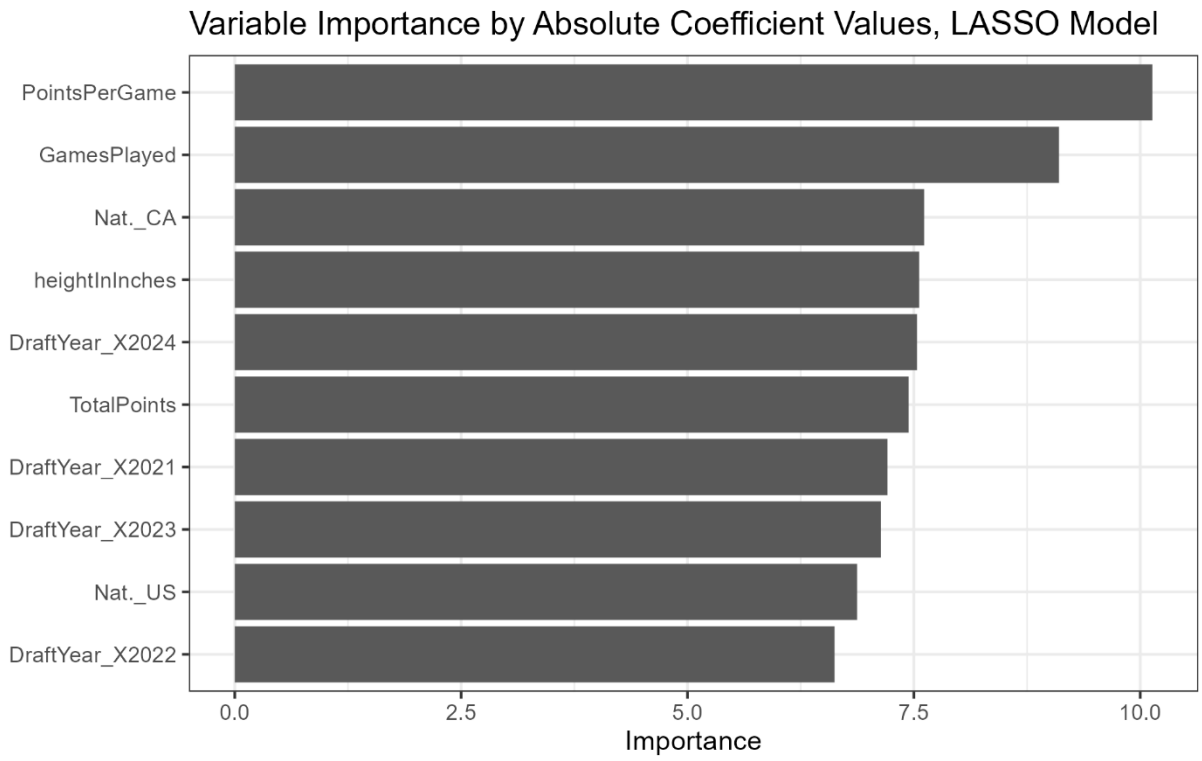

Figure 5.2 Variable importance for the LASSO model by coefficient size.

Points per game, games played, nationality, and height were all valuable in this model, with points per game having the largest variable importance by far. As expected, draft year is also an important variable in predicting expected goals. Logically, this makes sense because players who were drafted earlier in the dataset have generally played in more games and played in the NHL longer, leading to higher average flurry adjusted expected goals.

Ridge Regression

By nature of ridge regression, all 170 possible predictor variables (including dummy variables created from initial factor variables) were included in the model. Once again, ten-fold

cross validation was used to determine shrinkage penalty. Interestingly, the best shrinkage penalty was chosen to be $1 * e^{-10}$ (the penalty that minimized RMSE), a number so small that this hardly acts as any penalty, and essentially standard multiple linear regression is being performed. Perhaps even more interesting was that test RMSE decreased slightly from the LASSO model, changing to about 47.97. Below is a graph of test observations in terms of actual versus predicted values, illustrating a similar picture to the one seen from the LASSO model.

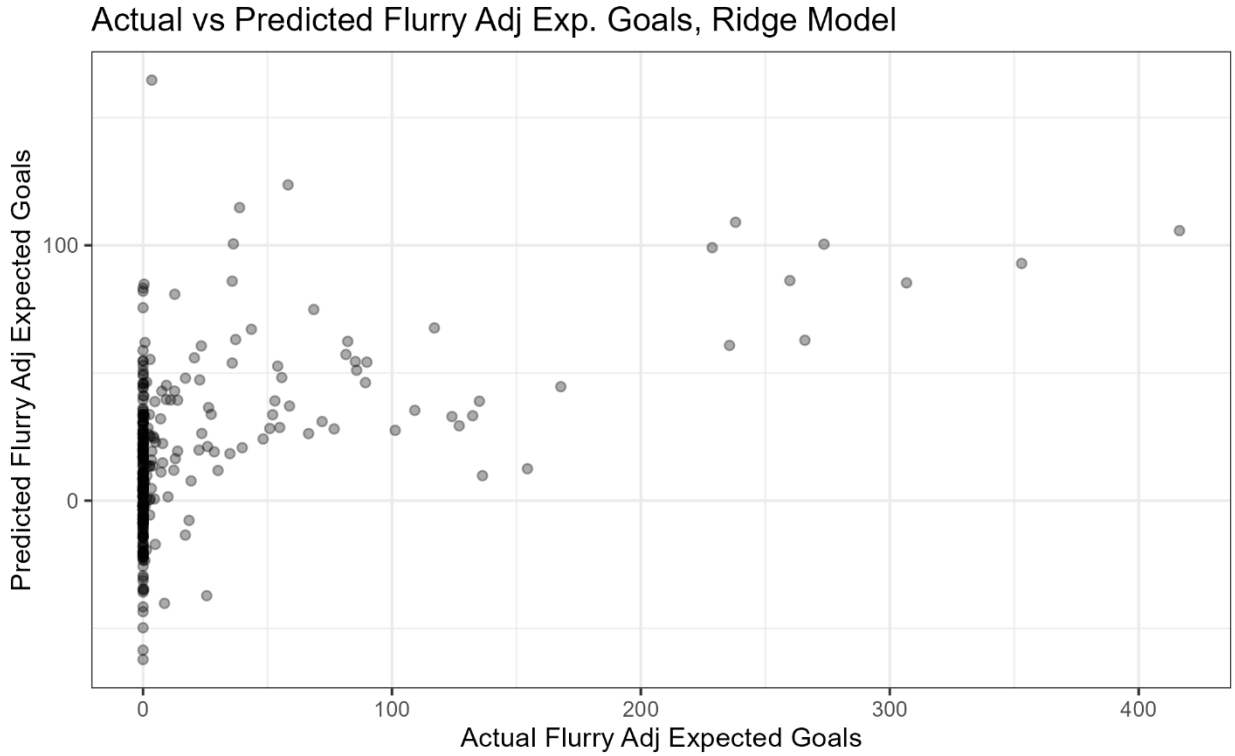

Figure 5.3 Actual vs Ridge Model Predicted Flurry Adjusted Expected Goals.

Once again, this scatterplot generates a similar takeaway as the LASSO model: an inability to interpret players with no expected goals, a lot of negative expected goal classification, and a significant underestimation of the few outlier players with actual expected goals above 100 (all players with more than 200 expected goals are all predicted to have similar

expected goal values close to 100). For comparison, let's examine the variable importance plot for this model by absolute value of coefficients.

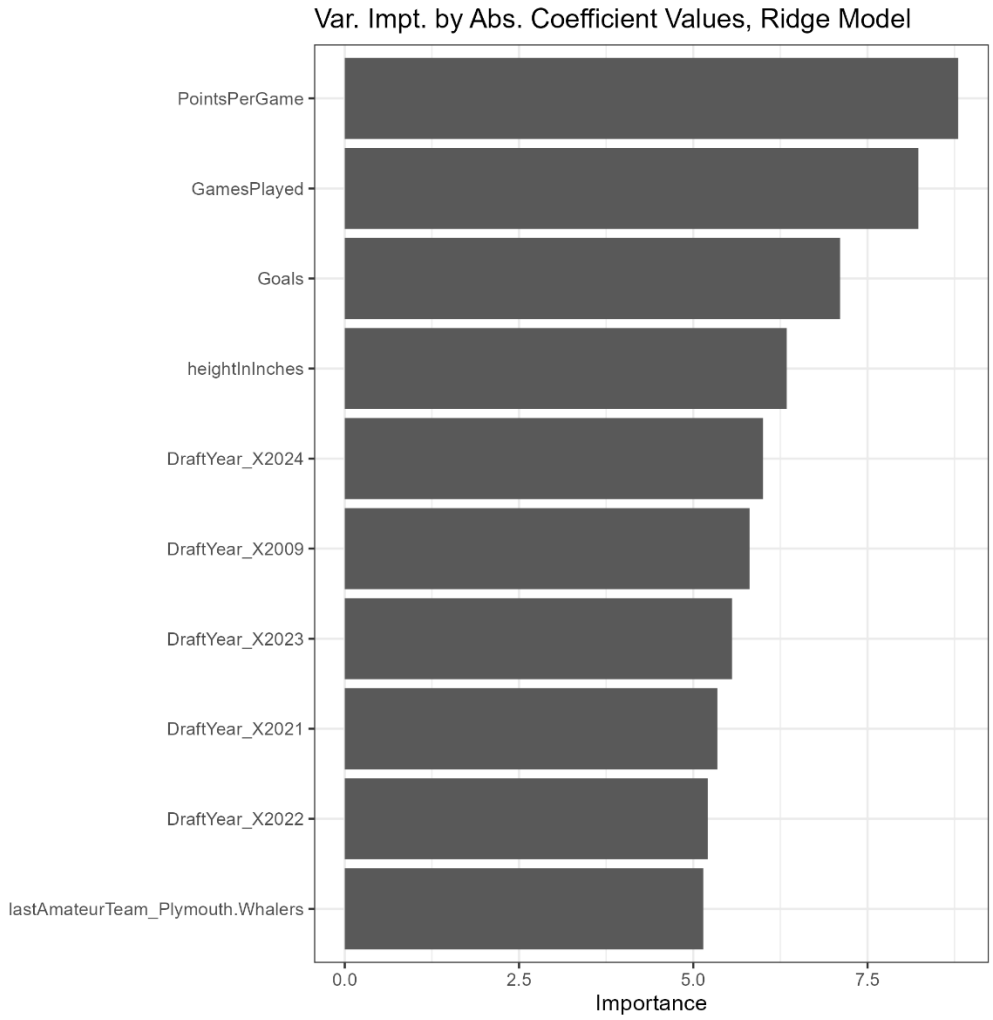

Figure 5.4 Ridge Regression Variable Importance by Coefficient Absolute Value.

Similarly to the LASSO model, amateur points per game and games played are extremely important predictors. Furthermore, even more draft years (2024, 2009, 2023, 2021, and 2022) are valued in the top ten most important predictors of expected goals, which makes sense logically by simply having differences in possible NHL career games played. Like the LASSO

model, this model was poor at predicting expected career goals, flurry adjusted, revealing that this subset of most important predictor variables was certainly not optimal for predictive ability.

Mixed Regression

Although it is possible to tune the mixture of LASSO and Ridge regression constraints used, I settled on a perfect 50-50 split, to investigate model performance with a balanced implementation of both methods. Still using ten-fold cross-validation to pick an optimal penalty and budget constraints based on minimizing RMSE, the optimal penalty was chosen to be 1 (as it was in the pure LASSO model). With this mixture of shrinking coefficients and eliminating some (reducing to zero), 115 out of 170 possible predictors were incorporated into this model (more than pure LASSO and fewer than pure Ridge, which makes sense due to the 50-50 split). Test RMSE remained very similar at 47.84 (better than the pure LASSO model and pure Ridge model by a very slim margin), and, as expected, the distribution of actual to estimated flurry adjusted career expected goals remained about the same.

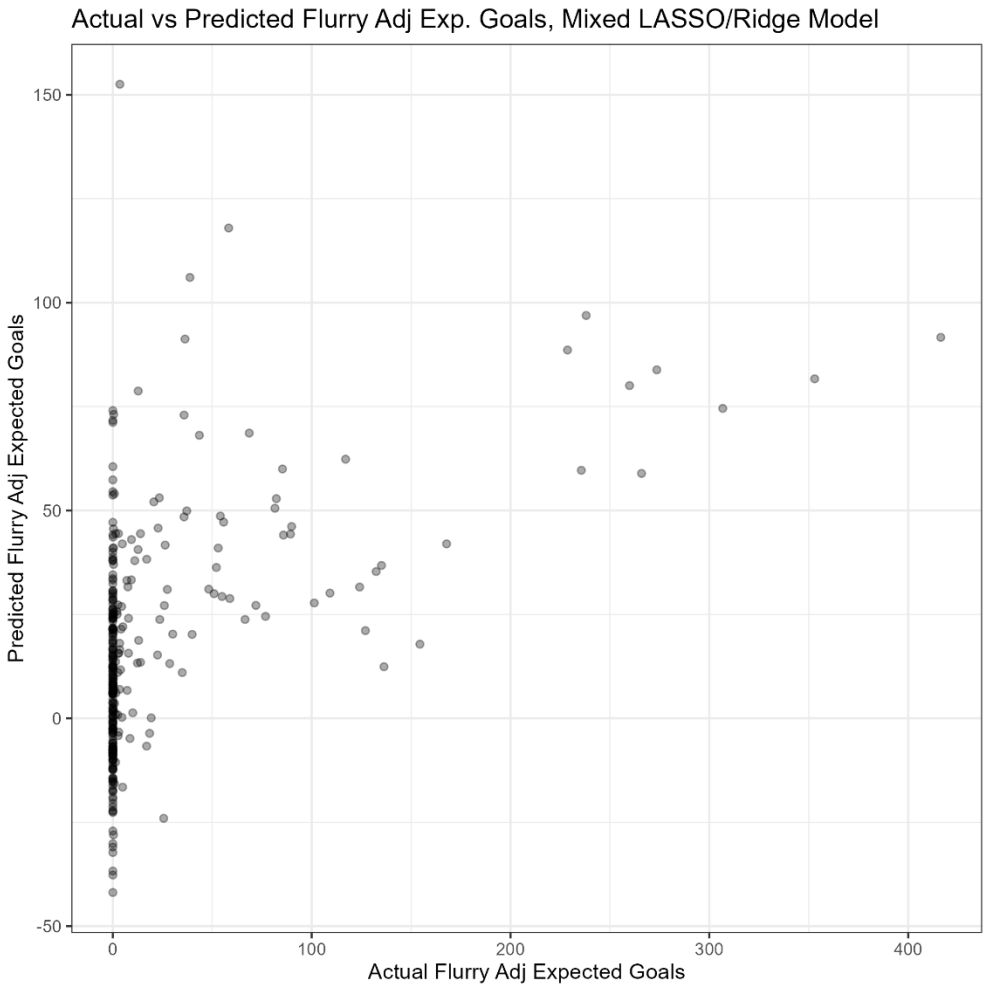

Figure 5.5 50-50 Mixed Model (LASSO/Ridge) Actual vs Predicted Flurry Adj. Expected Goals

For the third time, predictive accuracy for players with zero career expected goals is extremely low, as players in this category are classified anywhere from nearly negative 50 flurry adjusted expected goals to nearly positive 75. The estimate of flurry adjusted expected goals for some of the outlier players is still significantly lower than their actual statistics, albeit not as poor as these estimates for pure LASSO and Ridge Regression. The variable importance plot for the mixed model can be studied, once again plotting variables by coefficient size.

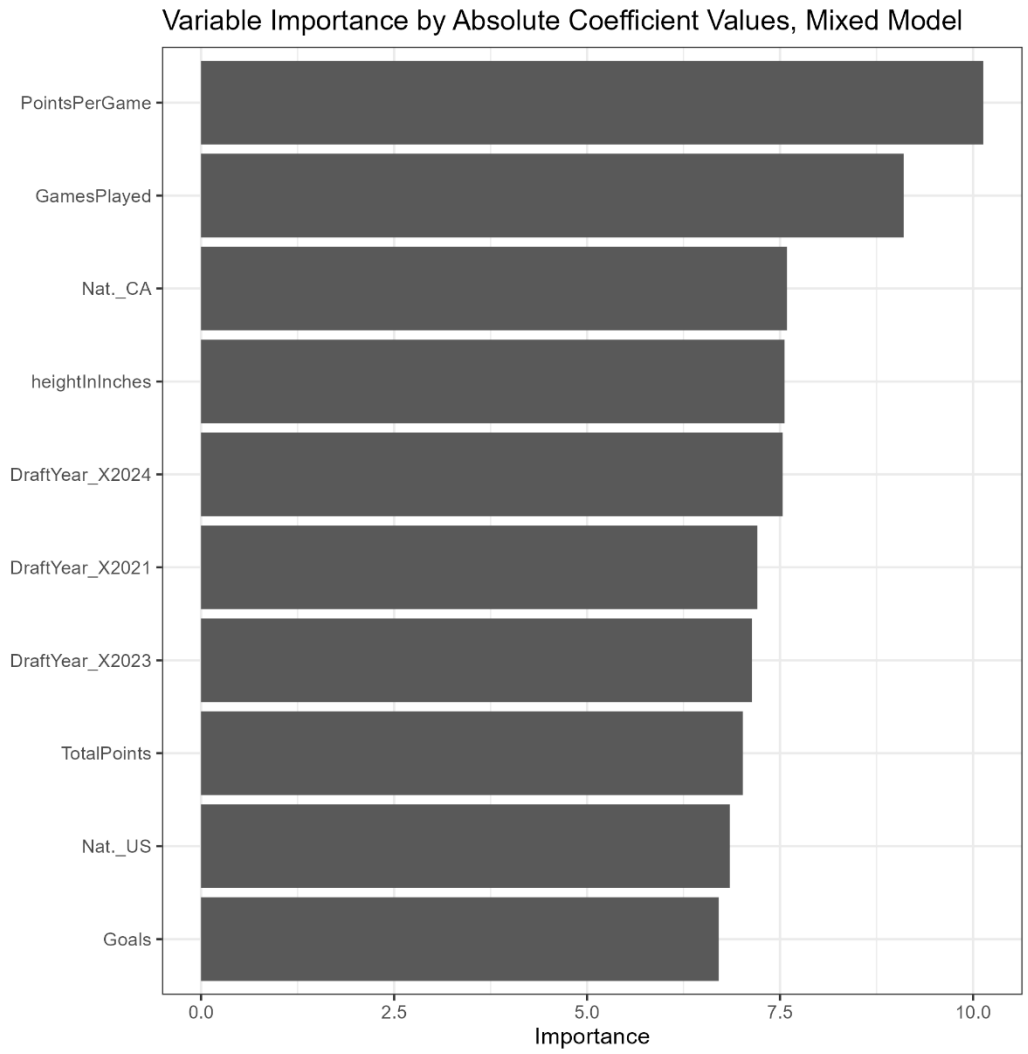

Figure 5.6 50-50 LASSO/Ridge Mixed Model Variable Importance by Coefficient Size.

Similar trends stick out from this model as well as the other linear methods: amateur points per game are highly valued for flurry adjusted career expected goals, as well as games played, overall goals, total points, nationality, and height. Again, draft year is shown to be an important predictor, which is logical based on the inclusion of many draft years in the dataset. Unfortunately, just like the pure LASSO and Ridge regression models, predictive capabilities were poor. Still, the hope was that tree-based modeling would predict at least a little bit better at the expense of some ease of interpretation.

Tree-Based Models

Boosted Trees

Unlike linear regression, using cross-validation on this model was particularly time-expensive, because of ten-fold cross-validation being utilized to tune learning rate (how quickly the model can use previous trees to improve) as well as tree depth, to find optimal values for each based on minimizing RMSE. The optimal parameter values found were a tree depth of 1 and a learning rate of 0.002, which is relatively small given usual learning rates (typically, a learning rate of about 0.1 is chosen, with 0.001 to 0.3 being the typical range of plausible values). Overall RMSE on test dataset was about 47.22, which is only better than the mixed model above by 0.62. For a little more than ten minutes of runtime, there was hope that the extra time might yield better predictions on the test set, but as the similar RMSE value indicated, those desires were not fulfilled.

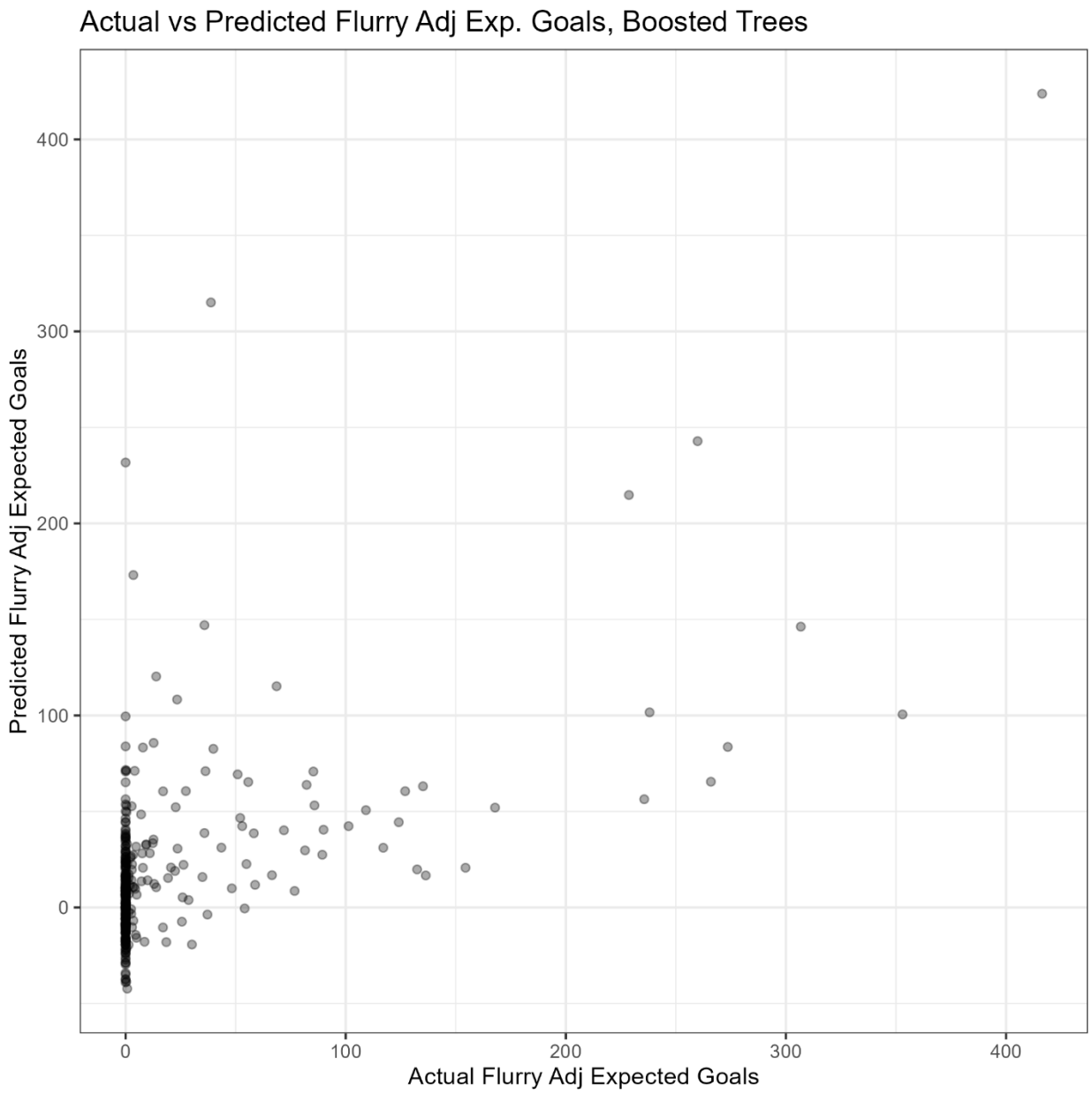

Figure 5.7 Boosted Trees Actual vs Predicted Flurry Adjusted Expected Goals.

Still, the boosted model was having significant difficulty classifying players with zero expected goals, as many players were still classified as having negative expected goals (which is obviously impossible). One notable improvement, though, between this boosted tree model and the linear models was with the outlier players (who had large flurry adjusted expected goal

values) were underestimated much less severely. Rather than flattening out at around a maximum estimation of 100 flurry adjusted career expected goals, the model predicted these players would have larger values than 100, even nearly predicting Steven Stamkos' 400 plus flurry adjusted expected goals very close to correctly. While there is still not much of an overall improvement in RMSE, this is certainly a positive note and refreshing to see a change in the pattern of consistently significant underestimation. When examining variable importance in this model, an additional refreshing change can be noticed.

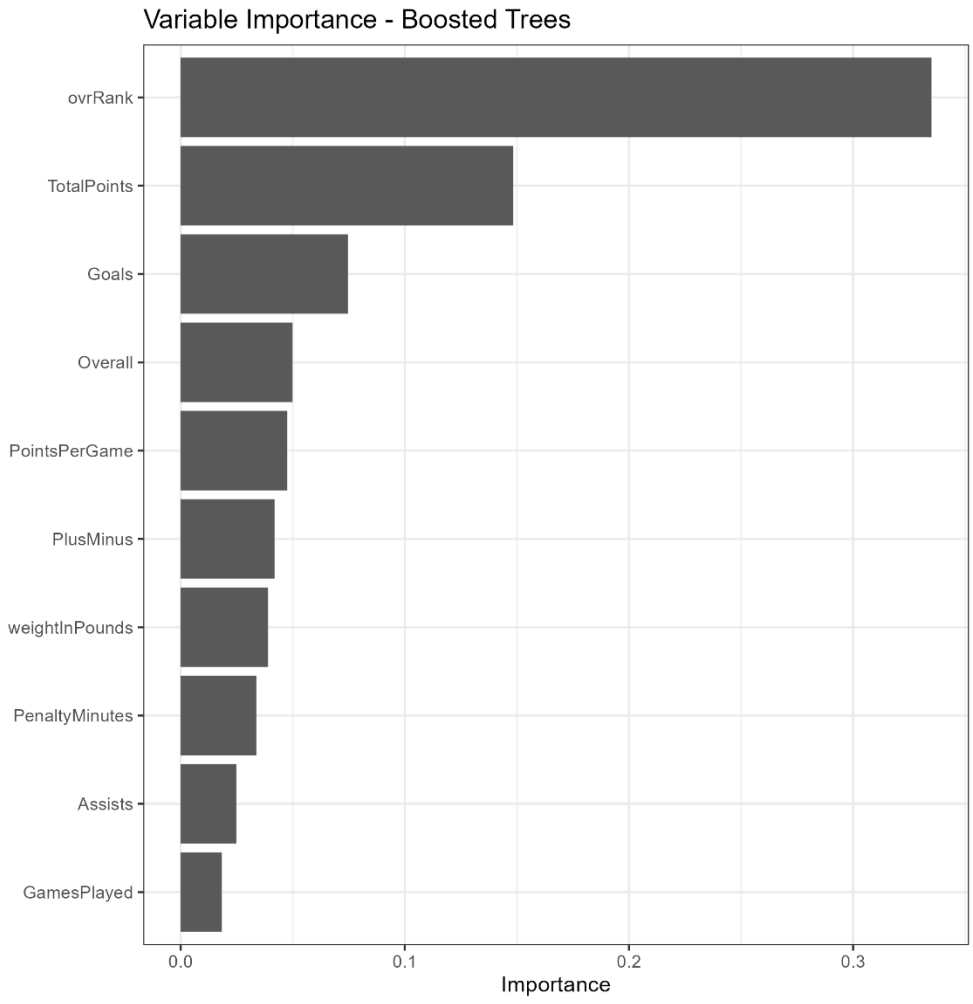

Figure 5.8 Variable Importance for the Boosted Tree Model.

Unlike in linear modeling, the draft year variables have dropped out of the top ten most important. With that said, it is important to note that the metric of variable importance has changed for decision trees (because there obviously are not coefficients to track the size of). While none of the linear models had overall player rank as a top ten variable by coefficient size, the boosted model has "ovrRank" as the far and away most important predictor of flurry adjusted career expected goals. Additionally, almost all the other top ten predictors in this model are related to amateur statistics, besides actual draft position and weight. Overall, despite the minimal gain in predictive accuracy, it was extremely encouraging to see the boosted tree model making use of variables that seem more likely to be related to actual flurry adjusted expected career goals in the real world than draft year (which is worthless when trying to project player success for a single upcoming entry draft).

Bagged Forest

Another method with the potential to decrease test RMSE (and thus increase predictive accuracy) was using a bagged forest. The characteristic of bootstrapped, repeated training sampling on the 1000 trees fit for this model presented this possibility. Because of this bootstrapped sampling, cross validation was not necessary to perform for this model (no tunable parameters were desired). In fact, test RMSE was found to be about 39.23, a significant decrease from all the previously discussed models (which hovered around a test RMSE of 47 and 48). When we examine the predictions more closely, the reasons for this reduction become clear.

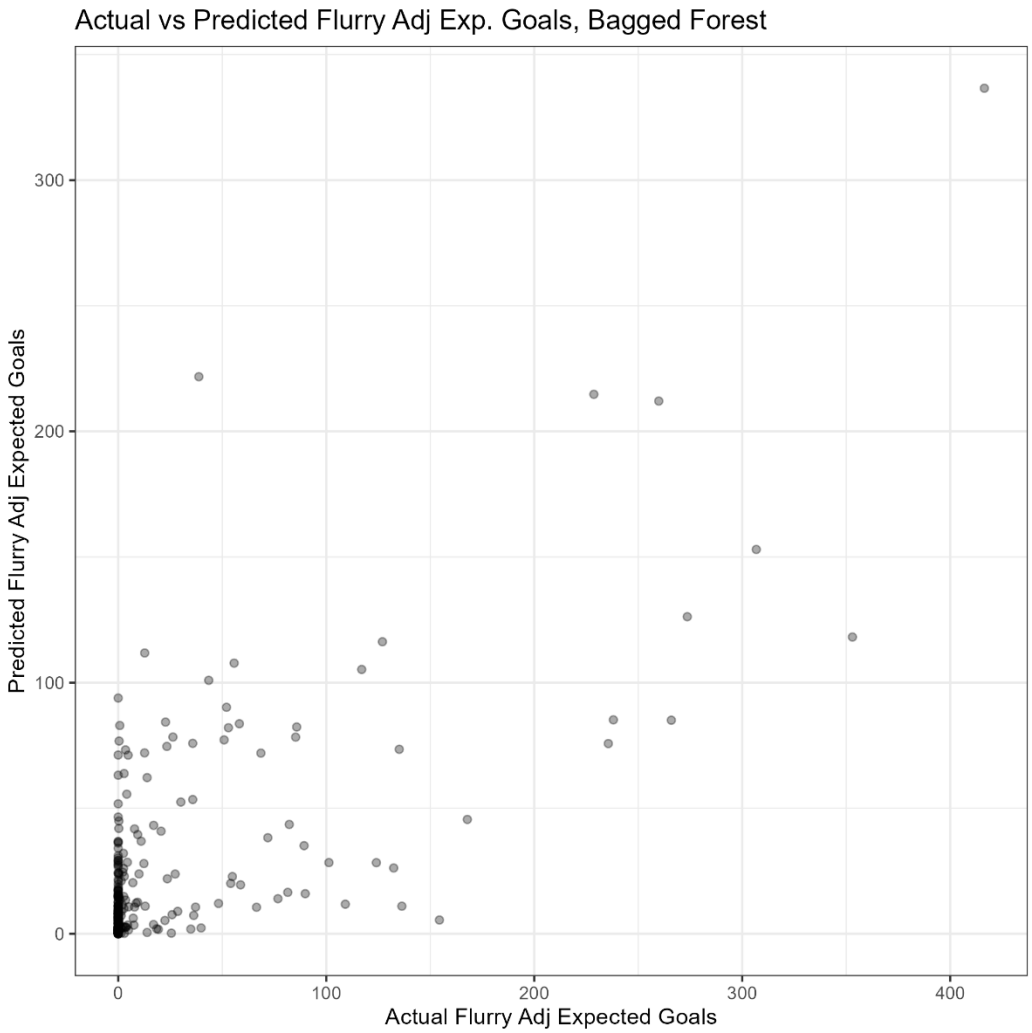

Figure 5.9 Bagged Forest Plot of Actual vs Predicted Flurry Adjusted Expected Goals.

Immediately, this actual versus estimated plot stands out as the most successful achieved so far. While there is still significant variation in the predicted flurry adjusted expected goals for players who had zero, this spread has decreased compared to any of the linear models and even boosted trees, with no players appearing to be predicted to have negative values. Further, prediction of the fewer players with large actual flurry adjusted expected goal values was significantly improved, which we can see by the relatively linear trend in points plotted. Additionally, examining variable importance for this model also reveals unique trends related to

model function. Note that variable importance, as chosen for trees in this analysis, measures how much training RMSE increases when the order of the variable splits in the trees is changed (called *permutation* importance). This metric is slightly more time-expensive to calculate, but the alternative metric, impurity, only measures raw reduction in variance created by each variable, so permutation importance felt like a more comprehensive metric to use in such a complex modeling task as this.

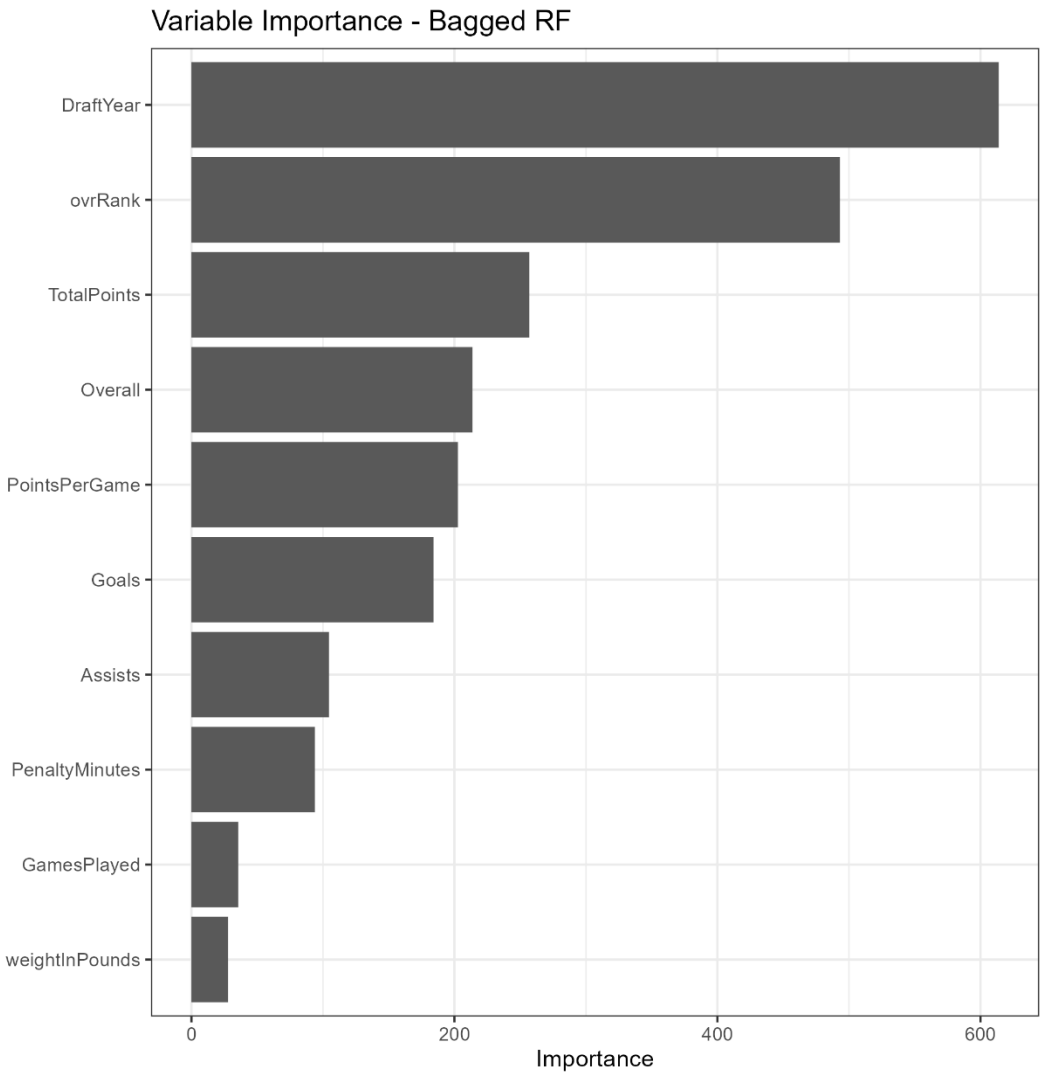

Figure 5.10 Permutation Variable Importance, Bagged Forest Model.

On the other hand, an unwelcome (but somewhat expected) trend that existed within the linear models has returned in this model: draft year as a primary predictor of career flurry adjusted expected goals. Beyond this predictor, which has the most variable importance by far, similar variables are contributing to model decisions as in the boosted tree, with overall rank moving from the most important predictor with boosted trees, to the second most important predictor here. Overall, nine of the top ten predictor variables found in this model are found in the boosted tree model, with the only difference being the inclusion of draft year here, compared to player plus/minus rating in the boosted tree model. Due to the apparent similarity in variable importance between the two models, the question of why draft year was included during bagged foresting and not significant in boosted tree learning is an interesting one. Nonetheless, the improvement in model performance with the entire test set is encouraging in bagged foresting, even though that improvement in accuracy was not extremely large.

Random Forest

Finally, random foresting was fit on this dataset in an effort to improve predictive accuracy even further. By the nature of this algorithm as an improved method of bagged foresting, there was hope that this approach might decrease test RMSE even further, thereby improving predictive capabilities. The only substantial change in code for this method was to limit the number of predictors the model can use at each tree split to mitigate correlation risk (once again, no cross-validation was needed due to the nature of bootstrapping repeated training samples). For this analysis, that limit was set as $m = \sqrt{p}$, where p is the total number of predictor variables in the model, and m is the subset of those predictors allowed to be used for tree splitting. While this parameter m is in fact tunable, attempting to tune the exact number of predictors the random forest was allowed to view was extremely cost expensive in code, thus

leading to using this conventional value instead. Overall, test RMSE was roughly 39.5, surprisingly indicating performance that was about the same as achieved in the bagged foresting model. We see evidence of these similarities in the actual versus predicted value plot below.

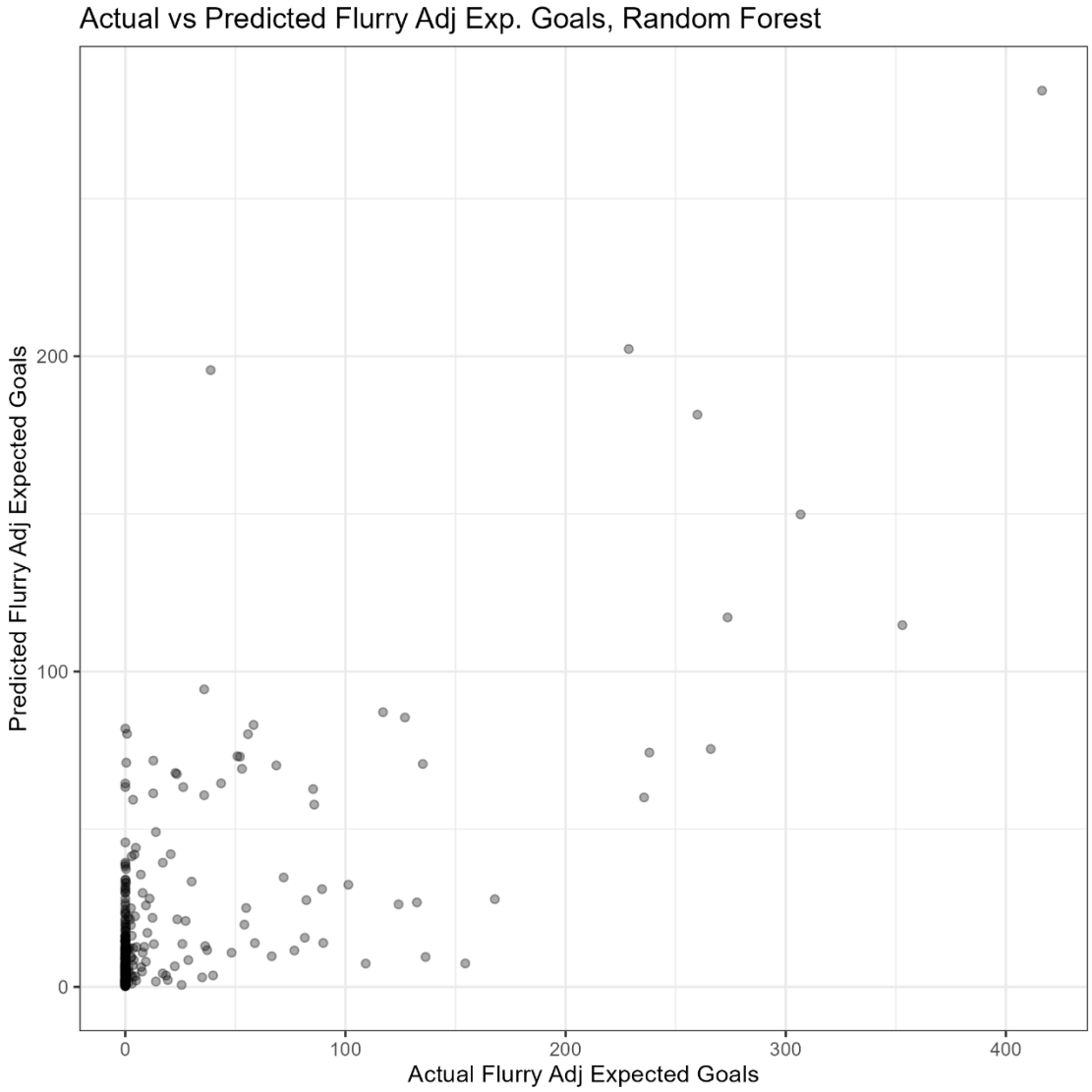

Figure 5.11 Actual vs Estimated Test Set Flurry Adj. Expected Goals with Random Foresting.

Similarly to the bagged foresting model, the problem of classifying players with negative expected goals has been eliminated, and overall spread in estimated values for players who had zero career flurry adjusted expected goals decreased slightly (albeit with a wide range of

prediction values remaining). With larger actual values, predictions were relatively similar to the bagged foresting model, still underestimating actual career flurry adjusted expected goals slightly, but with significant improvement from any of the linear models previously explored. To examine variable importance, we can once again view a plot of importance as the permutation value mentioned previously and compare.

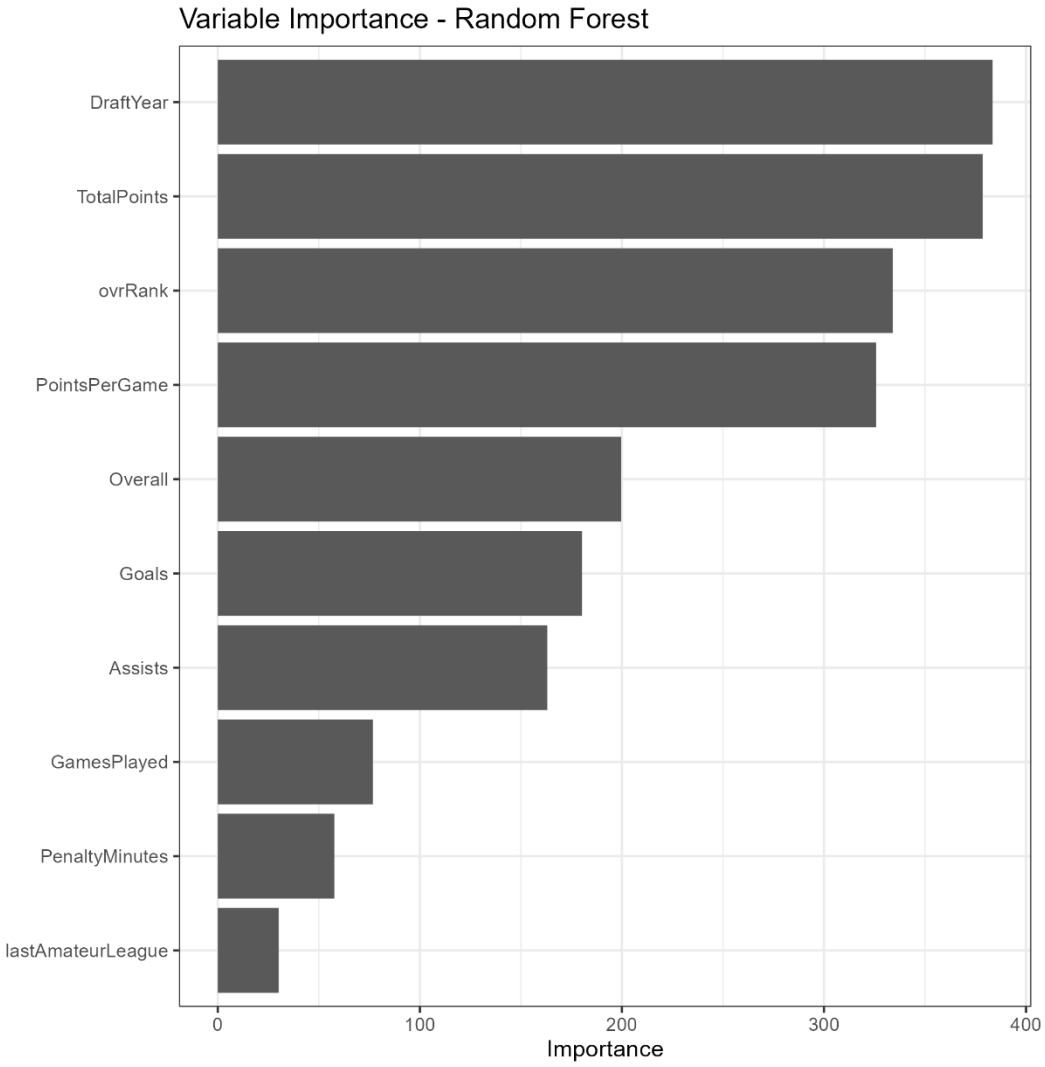

Figure 5.12 Permutation Variable Importance, Random Forest Model.

Similarly to the bagged forest method, draft year is the most important variable in expected goal prediction, with total amateur points the season prior to being drafted, overall player rank and amateur points per game right behind. Compared to the bagged forest model, it is encouraging to see these four variables close to the top, rather than draft year completely dominating variable importance on its own. As a whole, nine of the top ten variables were the same between bagged foresting and random forest, but the difference in order and magnitude indicated draft year being less important (draft year had an importance score of less than 400, compared to more than 600 in the bagged foresting model), which is a positive sign for model performance and ability to apply to it future prospect data.

Test Set Snapshot Evaluation

To provide a closer look at predictive accuracy for all models discussed, the test data was partitioned into six parts, based on the range of actual career expected goal values, so that each model could be analyzed in terms of how it performs on actual values spread throughout the dataset. As such, six players were chosen for this summary, randomly selected after partitioning the range of data. Below are their basic career backgrounds, obtained from Wikipedia:

1. **Paul Fischer (**Defensemen) – 138th overall pick in 2023. USNTDP player during the 2022-2023 season, and current NCAA hockey player for the University of Notre Dame. Actual flurry adjusted expected career goals: 0.
2. **Caleb Jones** (Defensemen) – 117th overall pick in 2015. USNTDP player during the 2014-2015 season. Currently playing for his fifth NHL team, after having split time in the American Hockey League (AHL, or direct minor league for the NHL) and NHL. Actual flurry adjusted expected career goals: 13.9.

3. **Zack Smith** (Center/Left Wing) – 79th overall pick in 2008. WHL player during the 2007-2008 season. Consistent depth player who spent 12 seasons at the NHL level. Actual flurry adjusted expected career goals: 101.3.
4. **Bo Horvat** (Center) – 9th overall pick in 2013. OHL player during the 2012-2013 season. Two-time NHL all-star who has served as both an alternate captain and captain at the NHL level. Actual flurry adjusted expected career goals: 235.62.
5. **Nathan MacKinnon** (Center) – 1st overall pick in 2013. QMJHL player during the 2012-2013 season. Seven-time NHL all-star, Stanley Cup Champion, and winner of several regular season awards like the Hart and Ted Lindsay trophies. MacKinnon is widely regarded as one of the best players in the NHL today. Actual flurry adjusted expected career goals: 306.75.
6. **Steven Stamkos** (Center/Wing) – 1st overall pick in 2008. Two-time Stanley Cup Champion, Seven-time NHL all-star, "Rocket" Richard Trophy winner (awarded to the NHL's leading goal scorer in a season). Actual flurry adjusted expected career goals: 416.75.

Each of these statistics is updated as of the end of the 2025 NHL season. Below is the table summarizing each model's predictions for each of the six players chosen to analyze.

PLAYER NAME	ACTUAL VALUE	LASSO EST.	RIDGE EST.	MIXED EST.	BOOSTED TREE	BAGGED FOREST	RAND. FOREST
STEVEN STAMKOS	416.4	81.26	105.75	91.65	423.77	334.44	284.2
ZACK SMITH	101.3	27.76	27.56	27.76	42.36	26.05	32.28
NATHAN MACKINNON	306.75	66.38	85.3	74.56	146.23	150.22	149.81
BO HORVAT	235.62	55.39	60.82	59.67	56.37	75.44	60.08
CALEB JONES	13.9	8.73	19.41	13.49	10.53	0.52	1.71
PAUL FISCHER	0	-12.91	-20.1	-15.22	-8.25	0.64	1.68

Table 5.1 Summary Predictions for Six Players from the Test Dataset by Model, versus Actual Career Flurry Adjusted Expected Goals.

This summary table highlights the trends in model prediction highlighted earlier: all three linear models significantly underestimated players with large flurry adjusted career expected goal values, while also predicting frequent negative values for Paul Fischer, a player with zero flurry adjusted expected goals. When moving into tree models, while accuracy did not make a monumental jump, underestimation of players with large values improved in all three models, while negative estimation of players with zero values decreased in addition.

Finally, the table below provides a summary of test set RMSE for each model.

MODEL	TEST ROOT MEAN SQUARED ERROR
LASSO	48.24
RIDGE REGRESSION	47.97
50-50 MIXED LASSO/RIDGE	47.84
BOOSTED TREES	47.22
BAGGED FOREST	39.23
RANDOM FOREST	39.5

Table 5.2 Summary of test RMSE for each model.

Overall, random forest and bagged forest produced the lowest test RMSE values of all models. Additionally, it is possible that optimal tuning of the exact number of variables the random forest could decrease test RMSE for this model even further.

Discussion and Limitations

Overall, projecting future player performance, particularly for an entire career, is a difficult task, and the performance of these models proved that. While perfect prediction is not possible, the overall estimates for each model left significant room for improvement. For the linear models specifically, these results certainly left the assumption of linearity (a linear relationship between predictor variables and the outcome) in question, particularly due to the prediction of negative flurry adjusted career expected goals, which is impossible given the nature of the statistic. This may speak to the inability for the models to decipher a clear pattern indicative of players with zero (or very few) expected goals, other than by considering draft year.

The discussion of draft year brings up an important point. The use of a response variable aggregated for an entire player's career certainly had its pros and cons. As a positive, it (in

theory) allows for projection of player's entire careers, rather than just one season or set of games. However, in a dataset with players drafted in different years, this presented a clear limitation in model performance, because two players of similar skill, with one being drafted earlier, will have different career expected goals, simply because one player has been in the NHL for a longer period and played in more games. One possible alternative to a statistic such as this would be to create rate statistics for each player (average their statistics into a per season or per game rate). This strategy may have performed better in this analysis, but it too contains some limitations that could have skewed the results. As an example, consider a player who gets injured early into a season, after playing in only five games. Suppose that this player was on a "hot streak" or playing significantly better than normal for those games, generating a high flurry adjusted expected goal statistic until getting injured. If rate statistics were calculated on a per game basis, this player would be interpreted as having an extremely high flurry adjusted expected goals average per game, for only those five games (it seems unlikely a player could keep generating scoring chances at such a high rate for an entire 82-game season). In this scenario, the rate statistic would be overestimating that player's actual impact. On the flip side, if rate statistics were calculated per season, this player would have an unfairly underestimated flurry adjusted expected goals value, because they only were able to participate in five games during that season (essentially undervalued due to an injury). As such, rate statistics can be flawed as well, potentially removing the influence of draft year as a predictor variable, at the expense of creating bias in the estimates of player contribution and success.

Another question might be about the choice of flurry adjusted expected goals to estimate player success. Of course, there is not a single "end all be all" metric that evaluates how successful a player is, in any sport. WAR in baseball (wins above replacement) has become

widely popular as a metric of player value, but similar attempts in hockey (such as point shares) have fallen short due to the subjective nature of their formulas. There are undoubtedly other metrics and statistics that are valuable to hockey player success, particularly for different skater positions (i.e., a defenseman might be more concerned with assists or plus minus rating than expected goals). However, flurry adjusted expected goals appeared to be the most objective, broadly applicable metric that could be measured on all skaters, and so it was chosen for this analysis.

With all these shortcomings in mind, the models still did show some predictability for players with high actual values, particularly for the tree models. Especially considering the nature of this data, with large proportion of zero values in the response due to very few players in late draft rounds spending significant time at the NHL level in their careers, this is a positive sign that these models demonstrated an ability to identify players with potential to have very successful careers, based solely on physical attributes, prospect rank, and a single season of amateur data.

Potential Future Research

As previously mentioned, these models could serve as a foundation for more advanced machine learning methods that incorporate both new prospect metrics and data. Firstly, it is reasonable to assume that significantly more data on player analytics is available to professional team scouts and those within the NHL, such as skating speed, shot velocity, and more. All these variables, if tracked at the amateur level, could be incorporated. Furthermore, more amateur level data could be collected and used to model career success, whether averaged by season or combined in some other way for each player. It might also be beneficial to consider differences

in amateur league rules, scoring trends, and other variations in play, because talent level and league-wide trends from the OHL are not perfectly applicable to trends in the ECHL (as an example). In the same vein, each model could be made more nuanced by performing machine learning on just one amateur league at a time, or by position (defenseman, center, or wing). Finally, further types of machine learning could be performed on this data, such as those used in the prior research section (K-means clustering, unsupervised learning, etc.) to identify patterns and trends that might predict player success.

All things considered, in an NHL hockey environment desiring new tools to aid player evaluation and career projection, this analysis provides a machine learning foundation that could be applied and expanded to project prospect success in the NHL for future draft classes, allowing teams to better examine what players they desire to take. By aiding prospect rankings and projections of career performance, NHL teams could improve overall draft success, particularly in later rounds.

Source Code

Access to R source code, as well as CSV files of data, can be found on this author's GitHub repository: https://github.com/aaroncg87/NHLProspectPrediction.git

References

"Bo Horvat." *Wikipedia*, Wikimedia Foundation, 15 Nov. 2025, en.wikipedia.org/wiki/Bo_Horvat.

"Caleb Jones (Ice Hockey)." *Wikipedia*, Wikimedia Foundation, 17 Oct. 2025, en.wikipedia.org/wiki/Caleb_Jones_(ice_hockey).

Deaner, Robert O., et al. “Born at the Wrong Time: Selection Bias in the NHL Draft.” 27 Feb. 2013, Accessed 13 Dec. 2025.

“Download Player and Team Data.” *MoneyPuck.Com -Download Data*, MoneyPuck, moneypuck.com/data.htm. Accessed 10 Dec. 2025.

“Elite Prospects - Hockey Players, Stats and Transactions.” *Elite Prospects - Hockey Players, Stats and Transactions*, Elite Prospects, www.eliteprospects.com/. Accessed 11 Dec. 2025.

Farah, Lou, and Joseph Baker. “Accuracy from the Slot: Evaluating Draft Selection in the National Hockey League.” *Scandinavian Journal of Medicine & Science in Sports*, 11 Nov. 2020, Accessed 13 Dec. 2025.

Farah, Lou, and Joseph Baker. “Eliminating Buyer’s Remorse: An Examination of the Sunk Cost Fallacy in the National Hockey League Draft.” *Scandinavian Journal of Medicine & Science in Sports*, 24 Feb. 2021, onlinelibrary-wiley-com.ezproxy2.library.colostate.edu/doi/pdf/10.1111/sms.13948.

“Jaro–Winkler Distance.” *Wikipedia*, Wikimedia Foundation, 9 Oct. 2025, en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance.

Koz, D., et al. “Accuracy of Professional Sports Drafts in Predicting Career Potential.” *Scandinavian Journal of Medicine & Science in Sports*, 3 Nov. 2011, Accessed 13 Dec. 2025.

Lou, Bryan C., et al. “Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury: An Analysis of 2322 Players from 2007 to 2017.” *Sage Journals*, 25 Sept. 2020, journals.sagepub.com/doi/10.1177/2325967120953404.

“Nathan MacKinnon.” *Wikipedia*, Wikimedia Foundation, 21 Nov. 2025, en.wikipedia.org/wiki/Nathan_MacKinnon.

“NHL Draft Prospects Rankings.” *Official Site of the National Hockey League*, NHL, www.nhl.com/draft/prospects. Accessed 10 Dec. 2025.

“NHL Entry and Amateur Draft History.” *Hockey-Reference*, www.hockey-reference.com/draft/. Accessed 11 Dec. 2025.

“Paul Fischer.” *Stats, Contract, Salary & More*, Elite Prospects, www.eliteprospects.com/player/637620/paul-fischer. Accessed 10 Dec. 2025.

Rago, Vincenzo, et al. “Identifying Key Training Load and Intensity Indicators in Ice Hockey Using Unsupervised Machine Learning: Research Quarterly for Exercise and Sport: Vol 96 , No 1 - Get Access.” *Taylor & Francis Online Journals*, 17 May 2024, www.tandfonline.com/doi/full/10.1080/02701367.2024.2360162.

Sharma, Abhishek. “Cross Validation in Machine Learning.” *GeeksforGeeks*, 29 Oct. 2025, www.geeksforgeeks.org/machine-learning/cross-validation-machine-learning/.

“Steven Stamkos.” *Wikipedia*, Wikimedia Foundation, 3 Dec. 2025, en.wikipedia.org/wiki/Steven_Stamkos#.

Tanner, Peter. “Flurry Adjusted Expected Goals.” *Moneypuck.Com -about and How It Works*, MoneyPuck, Jan. 2025, moneypuck.com/about.htm.

"USA Hockey National Team Development Program." *Wikipedia*, Wikimedia Foundation, 21 Nov. 2025, en.wikipedia.org/wiki/USA_Hockey_National_Team_Development_Program#cite_note-mediaguide-1.

"Zack Smith." *Wikipedia*, Wikimedia Foundation, 23 Oct. 2025, en.wikipedia.org/wiki/Zack_Smith.