Cap Hit Estimation Model for Players in the National Hockey League

Scott Jue and Laken Rivet

MSDS 456: Sports Performance Analytics

March 7, 2023

Image source: www.nhltraderumor.com/nhl-salaries, images.app.goo.gl/zhUTPfnFo9R8ofz97

Introduction

One of the most difficult decisions a front office in the National Hockey League (NHL) has to make is how to value the players in their organization. Valuation is an especially pressing issue in the league as a strict $82.5 million salary cap is enforced for all teams (Schram 2022). With restricted resources for signing players, teams need to be decisive in the offers they provide to prospects. The NHL is currently entering into new territory regarding the valuation of players. For the first time in four years, the salary cap was increased by $1 million between the 2021-2022 and 2022-2023 seasons. The cap is projected to continue increasing on a season-by-season basis for the next four years with variation in the amount of increase from $1 million to $4 million (Schram 2022). More cap space allows teams to devote more resources to their current players as well as seek out new talent. How to allocate these additional resources efficiently is the problem we aim to solve.

Our goal for this project was to create a machine learning model that can consider player information such as demographic data and past season performance to estimate an ideal annual cap hit to be offered upon signing. To do so, we built a number of different types of models and compared their performance on two datasets, one consisting of the active forwards in the NHL and one consisting of the active defensemen in the NHL. With this information, we hope to provide the front office of organizations an invaluable tool that will help facilitate data-driven decision making when it comes to valuing players.

Throughout the remainder of this report, we will first provide a literature review of similar projects and outline our methodology in determining the ideal type of machine learning model to create. From there we will discuss the data utilized, decisions made about how to clean and prepare the data for usage in the model, as well as justification for said decisions. Once the approach has been outlined, results of models will be presented visually and textually. To conclude the report we will provide a final recommendation of the ideal model for each type of player based on a comparison of the results and logistic considerations for an organization.

Literature Review

Increasing attention has been paid to the salary cap and its impact on talent acquisition in the NHL in recent years. As such, researchers have begun to utilize modern data analysis tools such as machine learning methods to evaluate team's effectiveness in the allocation of cap space as well as predict how to best allocate cap space in future signings. We have identified three previous studies with subjects and methods similar to our project that have provided insight into how best to approach solving this problem, and will present them in chronological order.

The first study, by Nugent (2018), utilized a random forest regression model to evaluate player salary data. He identified the top five most important features to salary prediction in his model to be draft year, birth year, time on ice, games played, and shots for. Further, he was able to explain 64.75% of variance in the data set with his model, a respectable result. Nugent (2018) demonstrated the value of using a random forest regression model, and we were intrigued to attempt to create one ourselves after seeing his results. However, there were some aspects of his model that we wanted to alter for ours. For example, he input 51 variables into his model, a number that we think is far too high to provide understandable results. Such alterations were kept in mind as we created our random forest regression model.

The second study was performed by Généreux and Xu (2021) and was focused on determining fair player salaries and evaluating general manager's effectiveness in negotiating player salaries. They estimated ideal salaries by utilizing a clustering method that grouped players by goal-scoring ability, physicality, position, and career stage. As such they were able to calculate the mean salary of the clusters and evaluate if a player was under- or overpaid by comparing their salary to the mean of their respective cluster. From this study, we noted the significant difference in salaries based on position and thus decided to model forwards and defensemen separately. While the clustering method did provide insights into player compensation, we felt there were more effective ways to predict an ideal salary offer, like the random forest regression model previously described, and decided not to pursue that methodology.

Jensen and Warren (2022) performed the final study considered. Their study was unique in that they classified cap hit, rather than annual salary, as their response variable. They created six different types of predictive models and compared their resulting root mean square errors (RMSEs) to determine the most accurate. They found that a cubist model performed the best followed closely by a random forest model. Following our examination of Jensen and Warren's (2022) work, we felt the best approach would be to create multiple models and identify the most effective rather than creating a singular model and attempting to improve it further. Additionally, we decided to use cap hit as our indicator variable as well as this amount is directed related to the salary cap compared to annual salary.

As random forest models were utilized in two of the three studies considered, we did some further research into why they were an appropriate model to use. A random forest is a tree-based model, which is particularly useful when there may be many interactions between predictor variables (Severini 2020). In essence, the random forest model consists of a specified number of decision trees that are fit to a random sample of the data set using a random sample of predictors. The result of the decision trees are then averaged to create an often quite accurate prediction (Beheshti 2022). There is much flexibility in the design of random forest models such as the number of trees, the depth of the trees, and the criterion to determine outcomes. In combination, the model's ability to handle predictor interaction, accuracy, and flexibility make it an ideal choice for this project.

Methods

Data Acquisition, Cleaning, and Filtering

Player data was acquired from www.capfriendly.com. Two separate datasets were obtained, one including forwards and one including defensemen. Identical data was collected for both positions which included variables with demographic data such as age, handedness, and position, variables with general skater statistics such as games played, points, shooting percentage, and time on ice, as well variables with advanced skating statistics like individual expected goals for, Corsi, and Fenwick. Prior to scraping the data, it was filtered for standard contracts only in order to prevent entry-level contracts (ELCs) from being included in the datasets. The NHL is unique in that it forces any player under 25 who is entering their first year in the league to sign an ELC (Murphy 2022). ELCs are capped at $925,000 annually, and vary in length from one to three years depending on the player's age. Thus, for younger players, their compensation does not match performance in the same way that players who have signed a standard contract does. With this in mind, we did not include players with ELCs in the datasets.

Once the data was acquired, cleaning was required for several variables. The data cleaning process was identical for both datasets. First, salaries and cap hits were recorded using commas and dollar signs (i.e. $12,00,000). To convert these entries to a usable format for the model, the commas and dollar signs were removed using string replacement commands and then converted to a numeric variable. The same was done for the shots for (SF) and shots against (SA) variables. Average time on ice (TOI) was recorded in a minutes and seconds format (i.e. 20:15). String manipulation was utilized again to separate the entries, with the colon as the delimiter, into seconds and minutes. The seconds were then converted to minutes and added to the minutes column to create the TOI_M variable, or average time on ice in minutes. Next, the variable Tot_MinPlayed which represents the total minutes played over the season by a player was created by multiplying the TOI_M by the number of games played (GP). Finally, rows that contained 'NA' values and duplicate rows were dropped from the data set.

Data filtering was also identical for the forward and defensemen datasets. In determining how to filter the datasets, our goal was to identify a minimum number of games played that allowed for the majority of the dataset to remain intact, but also required players to have played enough games so that their statistics were representative of their abilities. To do so, the proportion of players was plotted against games played and the cutoff for the bottom 20% of the datasets was identified to be 21 games for forwards and 19 games for defensemen (Fig. 1).

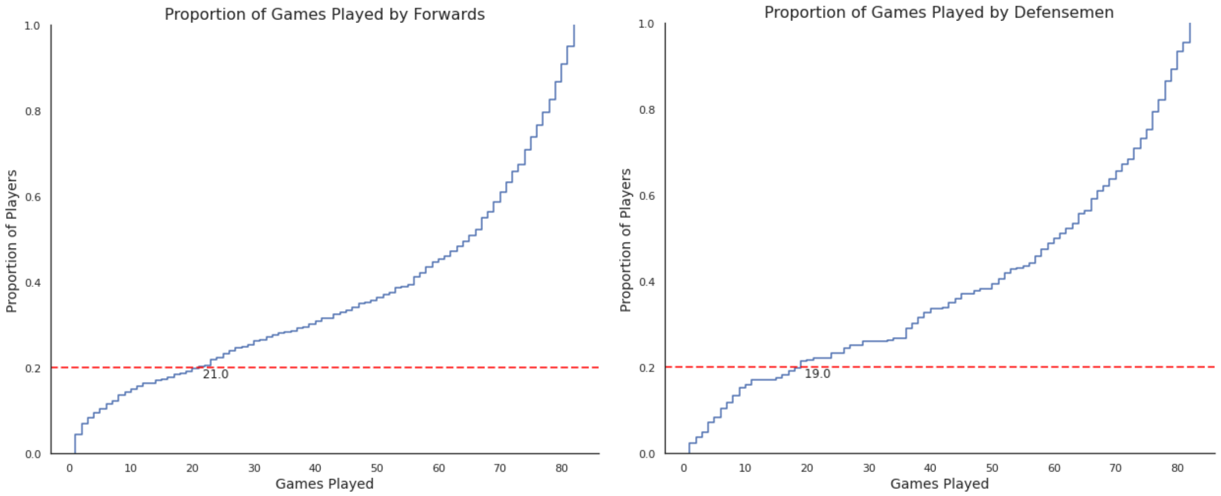
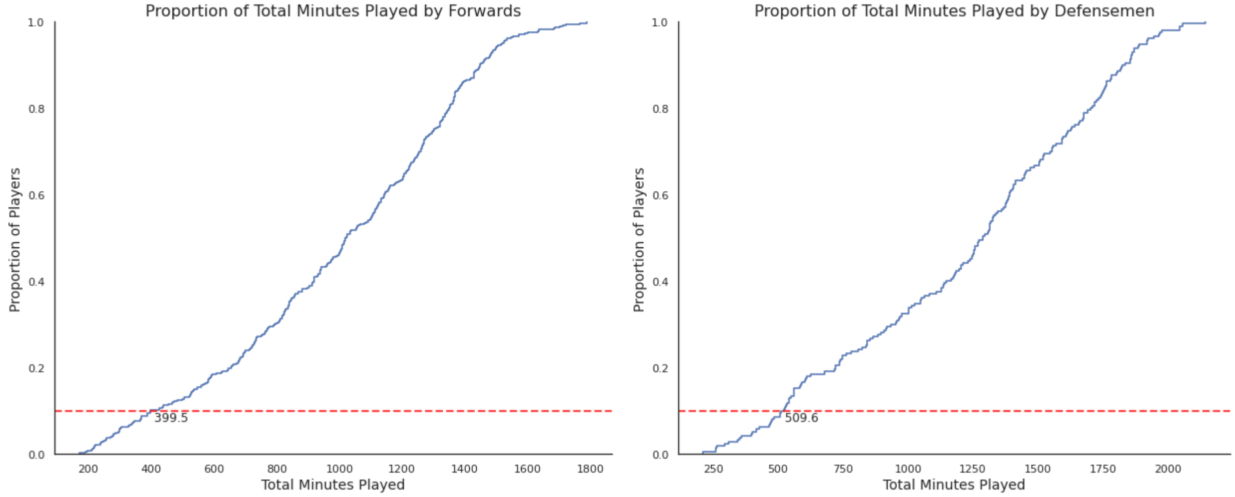


Figure 1. Proportion of players plotted against games played for forwards (left) and defensemen (right). The red dotted line represents the cutoff for the bottom 20% of players.

This process was repeated for total minutes played, except the bottom 10% was calculated rather than 20%. Cutoff lines were identified to be 399.5 minutes for forwards and 509.6 minutes for defensemen (Fig 2.).


Proportion of Total Minutes Played by Forwards

Proportion of Total Minutes Played by Defensemen

Figure 2. Proportion of players plotted against total minutes played for forwards (left) and defensemen (right). The red dotted line represents the cutoff for the bottom 10% of players.

The data was then filtered to include only players above the indicated cutoffs for both datasets which left entries for a total of 360 forwards and 189 defensemen.

The next step of data manipulation performed on both datasets was the creation of variables standardized by games played and total minutes played. We chose to standardize key performance metrics including goals, assists, shots, points, and plus/minus because we consider them to be related to the amount of time played. For example, a player who is included in the roster for every game of the season and averages upwards of 20 minutes on ice per game will have more opportunities to score compared to a player with less games played and less time on ice. As such, we found that a standardized statistic gives a better indication of how efficiently a player utilizes their time on ice. To create the first new variable, the aforementioned performance metrics were divided by games played. The same was done with total minutes played.

There were two categorical variables included in the data collected: handedness of the player and position. Many machine learning models, including regression models, cannot process categorical variables in their string format (Brownlee 2020). We used a technique called one-hot encoding to convert these categorical variables to binary variables, a format acceptable for use in regression and other types of models. Once the one-hot encoded variables were created, one from each category was dropped from the data sets to avoid redundancy in the models (Brownlee 2020).

Evaluating Relationships Between Variables

To understand the relationships between the variables, we looked at the correlations using a correlation matrix (Appendix A). From this matrix it was determined that handedness and position were not correlated to cap hit as pretty much all of the correlation coefficients were near zero. However, there were strong linear relationships with several performance related variables. The most notable correlation to cap hit was average points per game with a coefficient for forwards of 0.73 (Fig. 3). A similar, slightly stronger correlation was also observed in the defensemen data set with a coefficient of 0.75. Logically, it makes sense that the more points a player scores, the more value he provides to the team as this should ultimately increase the team's winning percentage. As such, the cap hit reflects the increase in value for more points produced in a linear pattern.

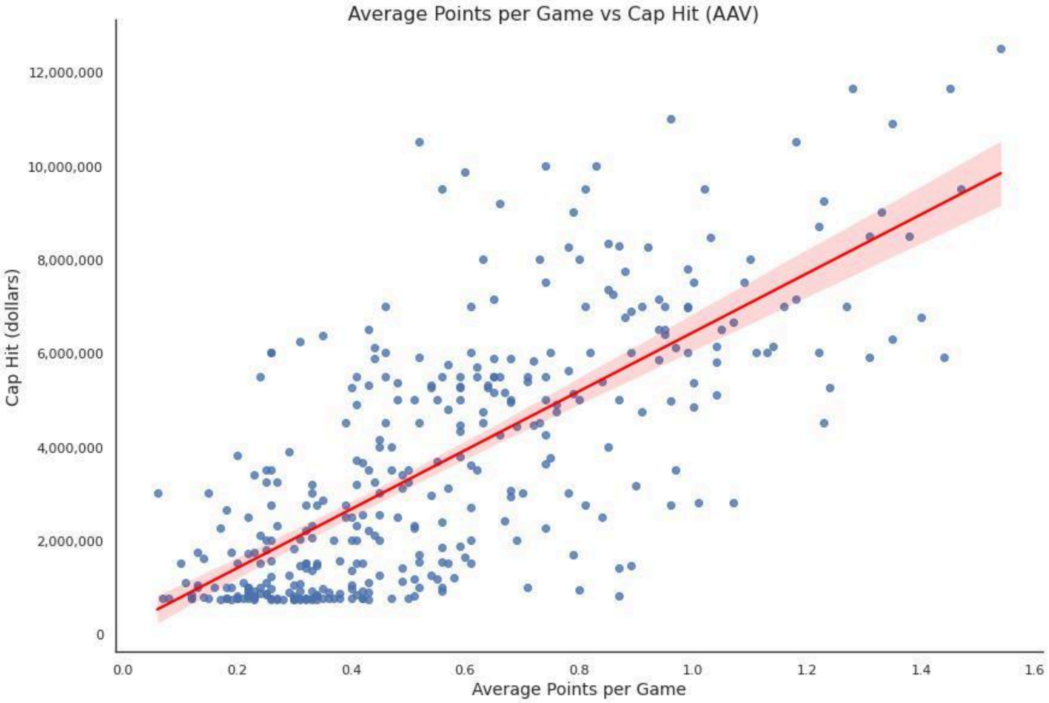


Figure 3. Average Points per Game vs Cap Hit (AAV) for forwards.

Another notable correlation observed was between cap hit and expected goals for (xGF), which is an estimate of the total goals a team is expected to score while a certain player is on the ice. The correlation coefficient was 0.70 for the forwards dataset (Fig. 4) and 0.77 for the defensemen dataset. As expected goals for is a metric that predicts goal scoring, it's logical that there is a strong, positive correlation with cap hit. Once again, if a player is performing well, they will be expected to continue performing well in exchange for being compensated well.

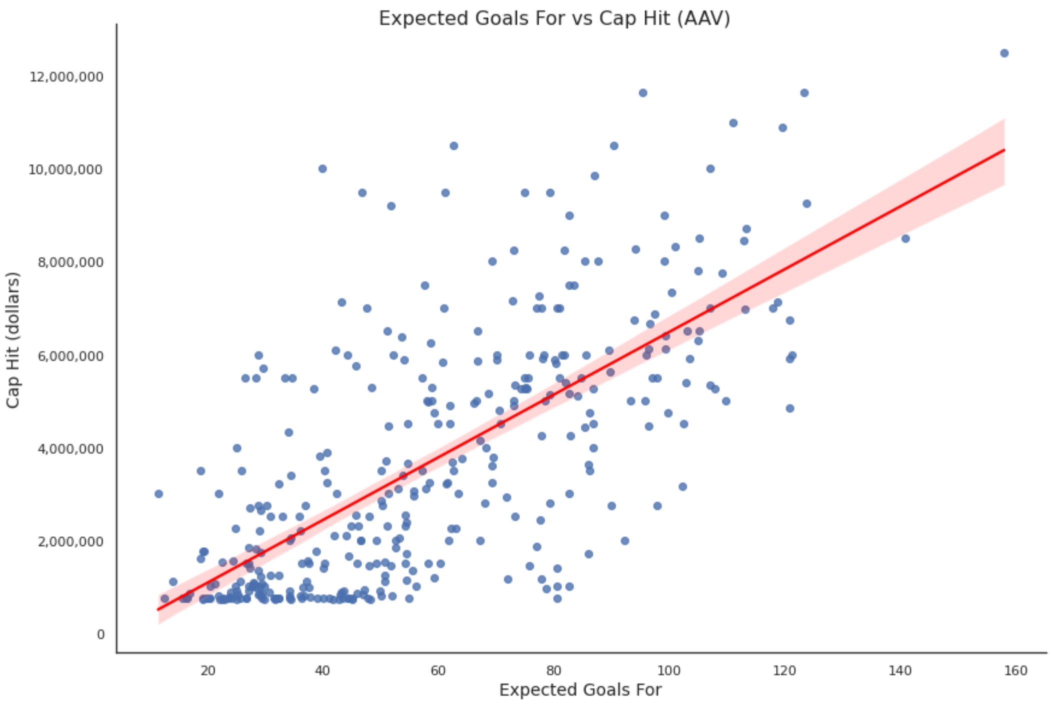


Figure 4. Expected Goals For vs Cap Hit (AAV) for forwards.

Model Types and Parameters

For model validation, we created a train/test split on the dataset using an 80/20 split prior to our modeling processing. This allows us to train our models on 80% of the data and then validate the model performances on remaining 20% unseen test data. In addition to providing a measure of accuracy on unseen data, this process can also help us determine if a model is overfitting the training data.

Upon identifying the linear nature of relationships between cap hit and several variables, the first type of model we decided to create was a multiple linear regression. The main benefit of utilizing this type of model is the interpretability of results, given the coefficients assigned to each predictor variable. Two methods of feature selection were utilized to determine the ideal set of predictor variables. First, we used the SelectKBest method with a regression function and instructed the algorithm to identify k = 10 features. As many of the variables identified displayed collinearity, we then utilized a variance inflation factor (VIF) feature selection technique to calculate VIF scores for each variable. The final set of variables was decided upon by comparing the top features from the SelectKBest method and their VIF scores, as well as our opinion on what variables would be conducive to an interpretable model. For forwards these variables included points per games played (P/GP), Corsi for (CF), and assists per minutes played (A/Min). For defensemen these variables included points per games played (P/GP), shots per games played (Sh/GP), and Corsi against (CA).

Two additional multiple linear regression models were created. The only difference between the original and subsequent models is the variables included. The first of the two included the variables selected by the original RF model, which will be discussed shortly. The second of the two has a reduced set of the same variables. A singular variable, Corsi for (CF), was dropped to create the reduced set as it had the lowest contributing weight and a negative coefficient value. It's not logical that a player's cap hit would be reduced as the total number of shots that occur while they're on the ice increases, so we decided to drop CF as a predictor.

Due to the strong collinearity observed amongst the variables in both datasets, we also created a ridge regression model. By including a penalty factor in the regression equation, a ridge model will prevent the collinearity from inflating the coefficients of the variables (Ashok 2022). The same variables were inputted into the ridge regression model as the original multiple linear regression model.

Next, as discussed in the literature review section, we created a random forest (RF) model. For the initial RF model, the default parameters of n = 100 trees, an unlimited depth of trees, and mean squared error as an accuracy metric were used. The model was fed all variables and automatically selected the variables it found to be most crucial to prediction. Three additional versions of the RF model were created. In the first, hyperparameter tuning using random search and grid search was performed to identify ideal parameters. The resulting parameters, which were used in the model, can be observed in Table 1. Again, the model was fed all variables and automatically selected which variables to include.

Parameter	**Forwards**	**Defensemen**
max_depth	11	11
max_features	'sqrt'	'log2'
min_samples_leaf	2	2
min_samples_split	2	12
n_estimators	100	700

Table 1. Resulting parameters from hyperparameter tuning of RF model.

We also explored performing dimensionality reduction on the dataset to reduce the number of input variables for our RF models. As such, the next version of the RF model utilized principal component analysis (PCA). As is required for PCA, all variables were first standardized using a standard scaler. Once the components were created, the number of components included in the model was determined using the calculated cumulative variance ratio. For both datasets, 6 components were used and were calculated to have cumulative variance ratios of 0.97. The default parameters of the RF model, as mentioned above, were used in this model as well.

The final version of the RF model created used both hyperparameter tuning and PCA. The same 6 principal components for both datasets, outlined in the previous paragraph, were fed to the model. The resulting parameters used, by position, can be found in Table 2.

Parameter	Forwards	Defensemen
max_depth	14	12
max_features	'log2'	'log2'
min_samples_leaf	2	2
min_samples_split	2	2
n_estimators	200	300

Table 2. Resulting parameters from hyperparameter tuning of RF model with PCA.

The final model type included in this analysis is a variant of a tree model called a gradient boosted (GB) machine. We chose to include this model as it varies slightly from a RF model in that a GB model builds one tree at a time and uses each new tree to correct errors in the previous tree (Ravanshad 2018). This error correction can help make GB models slightly more accurate with predictions compared to an RF model. In creating the GB model, the default parameters of n = 100 trees, an unlimited depth of trees, and squared error as the loss function to be optimized. As with the RF models, the GB model was fed all variables and automatically selected which variables to include.

Results

Forward Model

When analyzing the results and performance of the models on the test data, we looked at the RMSE and R-squared values to determine the best fitting model. Table 3 summarizes the results of our forward models. The random forest model had the best results, having the lowest RMSE and the highest R-squared value. This is consistent with previous research that also had found random forest models to be an appropriate fitting model for predicting salary and cap hit (Nugent 2018, Jensen and Warren 2022). However, random forest models are considered a black box method and as such, are more difficult to interpret than traditional linear regression models. For this reason, we took a hybrid approach and have selected the linear regression model using the reduced random forest variables as the predictor variables. The RMSE is only slightly higher than the random forest model and the R-squared is only slightly lower. This model is able to explain approximately 63% of the variance in cap hit for forwards and has a RMSE of approximately $1.65M. We believe this slight decrease in model performance is worth the trade-off for the improved interpretability of a linear regression model. The NHL is behind most other professional sports in its overall acceptance of analytics, but the sport is slowly evolving to include more analytics at a professional level (Goldman 2022). For this reason, we felt that it was more appropriate to opt for a more interpretable and explainable model.

Model Type	RMSE	R-Squared
Linear Regression	1,852,131.16	0.54
Ridge Regression	1,866,009.04	0.53
Random Forest	1,615,601.38	0.65
Gradient Boosted Machine	1,646,485.05	0.64
Linear Regression with RF Variables	1,648,004.34	0.64
Linear Regression with RF Variables (Reduced)	1,652,445.09	0.63
Random Forest (Hyperparameter Tuning)	1,685,450.55	0.62
Random Forest with Principal Component Analysis (PCA)	1,699,704.61	0.61
Random Forest with PCA (Hyperparameter Tuning)	1,730,874.01	0.60

Table 3. Forward Model Results Summary on Test Data.

The final forward model variables and coefficients can be seen in Table 4. We have determined that a player's points per games played (P/GP), assists per games played (A/GP), age, and Expected Goals For (xGF) are the best variables to use for predicting a player's cap hit. We can interpret the parameters of this model as for every 1 unit increase in P/GP the player's predicted cap hit will increase by approximately $2,150,000. For A/GP, every 1 unit increase will increase the predicted cap hit by approximately $3,328,000. For every 1 year increase in age, the predicted cap hit will increase by approximately $202,000, and every 1 unit increase in xGF will result in an approximately $27,500 increase in cap hit. The intercept value can only be interpreted within the context of this dataset as it is a negative value and having a negative cap hit is not possible. This value implies that a player's predicted cap hit is approximately -$5,713,500 if all variables are equal to 0. However, this is also not feasible since the age variable cannot be 0 since it has a minimum value of 18. As such, these values were derived to fit this specific dataset and should be only interpreted within the context of this dataset.

Coefficient	Value
Intercept	-5,713,540.94
P/GP	2,149,993.66

A/GP	3,328,038.44
Age	201,815.55
xGF	27,524.53

Table 4. Linear Regression (Reduced RF Var.) Forward Model Coefficients.

Evaluating Model Performance

In order to obtain meaningful inferences from the regression model, the residuals should be normally distributed. The residuals for the forward model show a normal distribution (Fig. 5), so this model does not violate the assumption of normality for regression models.

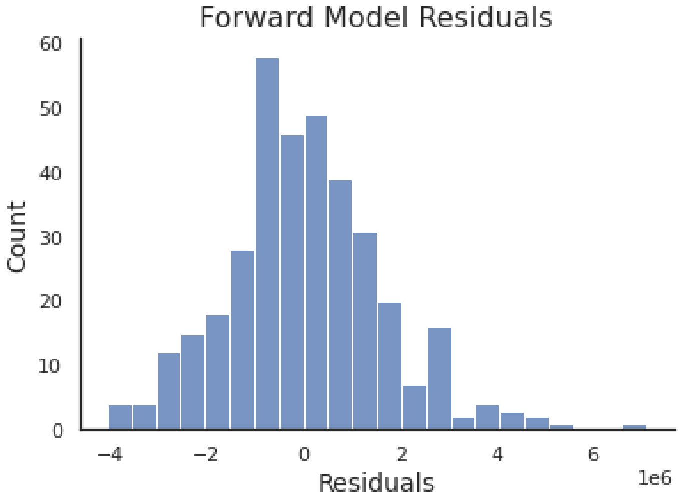


Figure 5. Histogram of Residuals for the Linear Regression (Reduced RF Var.) Forward Model.

Using the model to predict cap hit, we then compared the results to the actual cap hit and looked at the top 10 forwards with the highest residual error (Fig. 6 and Fig. 7).

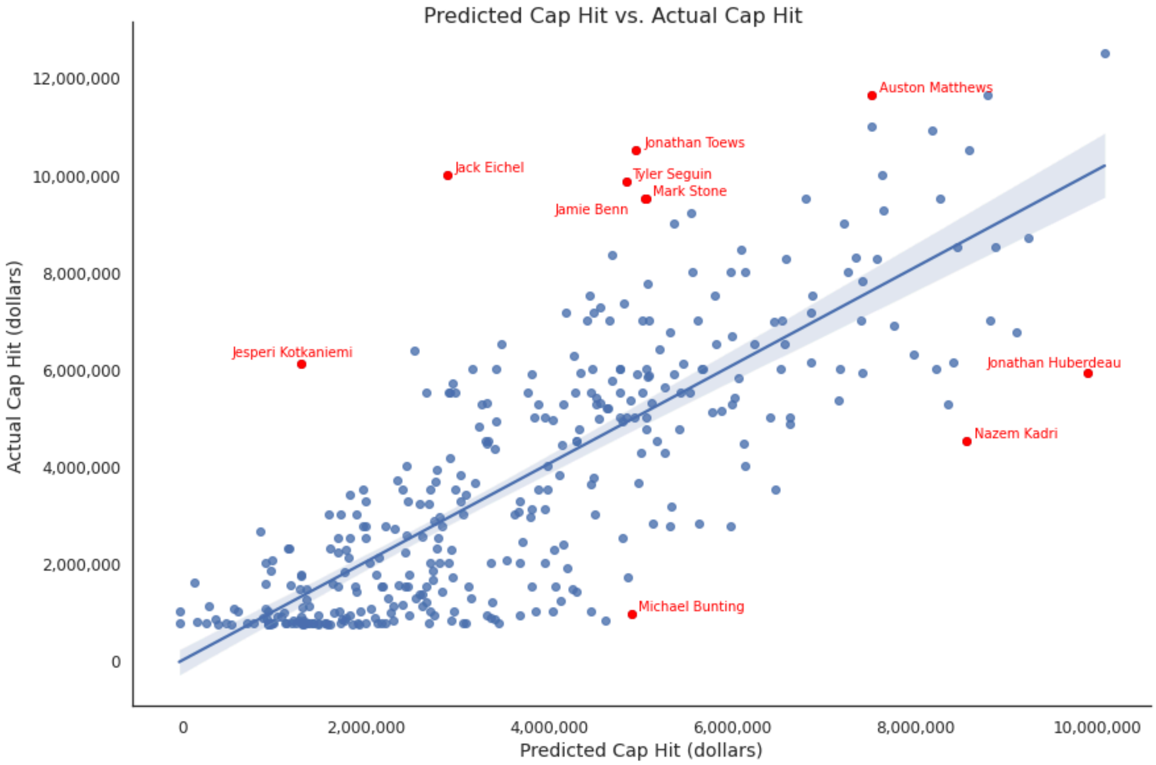


Figure 6. Predicted Cap Hit vs Actual Cap Hit for forwards.

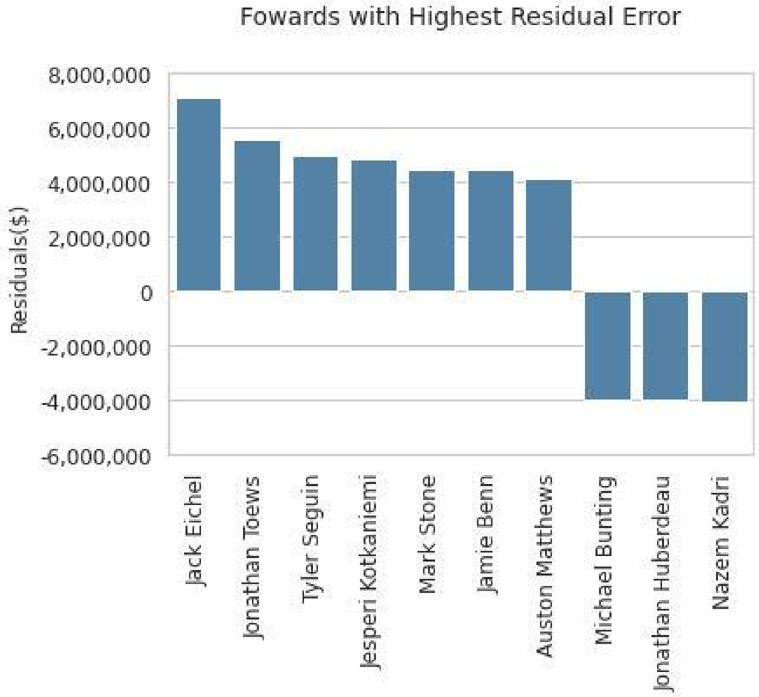


Figure 7. Forwards with Highest Residual Error.

Highest Under-predicted Forwards

There is some significant variance between the predicted and actual cap hit for forwards. We investigated a few of these players to determine why the model's predictions were so far off. First, looking at the player with the highest residual, Jack Eichel, we can see that the model predicted a cap hit of approximately $2.9M; however, his actual cap hit is $10M. Jack Eichel is in a unique situation in that he was traded to a new team (Vegas Golden Knights) and also coming off a disk replacement surgery in his neck which led him to miss the first half of the season (Webster 2021). He only had 34 games played, so his stats may have been negatively impacted by this injury and resulting surgery as well as trying to adapt to a new team.

Next, we investigated Jonathan Toews' high residual. Toews' predicted cap hit is approximately $5M while his actual cap hit is $10.5M. Toews was diagnosed with COVID in 2020 and missed the entire 2020-21 season. Although he did return to play in the 2021-22 season, he had lingering symptoms that likely affected his performance. Just recently, in February 2023, he decided to step away from hockey in order to focus on his health and fully recover from these lingering symptoms (Cohen 2023). Furthermore, Toews is also the captain of the Blackhawks, and has led the team to 3 Stanley Cup championships. Therefore, his leadership skills have proven to be valuable; however, the current performance data does not account for these types of variables. Therefore, we have determined that factors such as injuries, current health, leadership qualities are off-ice variables that impact a player's performance and cap hit, but are not currently accounted for in the model. As such, these are important limitations of this model since we have only trained the model using on-ice game stats and demographic data.

Another player we investigated was Tyler Seguin who has an actual cap hit of $9.85M but a predicted cap hit of approximately $4.85M. The actual cap hit comes from his 2018 8-year contract extension where he had his first and only 40+ goal season, consistently scoring 70+ points for consecutive seasons. Since that 2017-18 season, Seguin has only eclipsed 30+ goals once and 70+ points once. As such, his contract was valued during his prime and he has since yet to produce at the same level as he was in 2018. So, the model underpredicted his actual cap hit due to his drop off in play in recent years.

Highest Over-predicted Forwards

Alternatively, the model can also overpredict a player's cap hit. The players with the largest residual for overpredicted cap hit are Nazem Kadri, Jonathan Huberdeau, and Michael Bunting. Since Kadri had the highest overpredicted cap hit, we investigated Kadri's stats and current contract to see why his cap hit was predicted much higher than his actual cap hit. Kadri's 2021-22 actual cap hit of $4.5M was overpredicted at approximately $8.6M. In the dataset used for the model, Kadri was actually in the last year of his 6-year contract that he originally signed with the Toronto Maple Leafs back in 2016. He recently signed a new contract with the Calgary Flames this past season that has a cap hit of $7M. This new cap hit is more in line with his performance metrics. Additionally, his 2021-22 season stats were a bit of an anomaly as he had a total of 87 points which is the most he has had in his career. The second most points he has had prior to that season was 61 points in the 2016-17 season. So, another limitation to this model is that it only uses data from a single season. Player's who are having a down season are likely to have a predicted cap hit that is lower than their actual cap hit. On the other hand, players that are having an abnormally good season or a breakout season are likely to have a predicted cap hit that is higher than their actual cap hit.

Defensemen Model

Table 5 summarizes the RMSE and R-squared for each of the defensemen models tested. For defensemen, the best performing model was the ridge regression model using points per games played (P/GP), shots per games played (Sh/GP), and Corsi against (CA) as the predictor variables. This model was able to explain 73% of the variance observed in the cap hit amounts for defensemen using 2021-22 regular season data. While technically the linear regression model with the random forest variables was the best performing model, this model had several negative coefficients that went against conventional logic. Therefore, this model is not feasible and we selected the ridge regression model as our best performing and most interpretable model.

Model Type	**RMSE**	**R-Squared**
Linear Regression	1,489,827.30	0.72
Ridge Regression	1,469,651.93	0.73
Random Forest	1,741,142.05	0.62
Gradient Boosted Machine	1,753,149.31	0.62
Linear Regression with RF Variables	1,465,443.25	0.73
Linear Regression with RF Variables (Reduced)	1,606,625.64	0.68
Random Forest (Hyperparameter Tuning)	1,683,122.41	0.65
Random Forest with Principal Component Analysis (PCA)	1,977,605.09	0.51
Random Forest with PCA (Hyperparameter Tuning)	2,045,544.33	0.48

Table 5. Defensemen Model Results Summary on Test Data.

The defensemen model variables and coefficients can be seen in Table 6. We can interpret the parameters of this model as for every 1 unit increase in P/GP the player's predicted cap hit will increase by approximately $3,220,000. For Sh/GP, every 1 unit increase will increase the predicted cap hit by approximately $1,562,000. For every 1 unit increase in CA, the predicted cap hit will increase by approximately $1,800. The intercept value can only be interpreted within the context of this dataset as it is a negative value and having a negative cap hit is not possible.

This value implies that a player's predicted cap hit is approximately -$2,224,000 if all variables are equal to 0. However, this is also not feasible as a player with no points or shots while on ice would likely not be in the NHL. As such, these values were derived to fit this specific dataset and should be only interpreted within the context of this dataset.

Coefficient	Value
Intercept	-2,224,412.19
P/GP	3,219,495.02
Sh/GP	1,562,115.93
CA	1,803.19

Table 6. Defensemen Ridge Regression Model Coefficients

Evaluating Model Performance

The histogram of the residuals show that the residuals for the defensemen model approximate a normal distribution (Fig. 8). Similar to the forward model, this model also does not violate the assumption of normality for regression models.

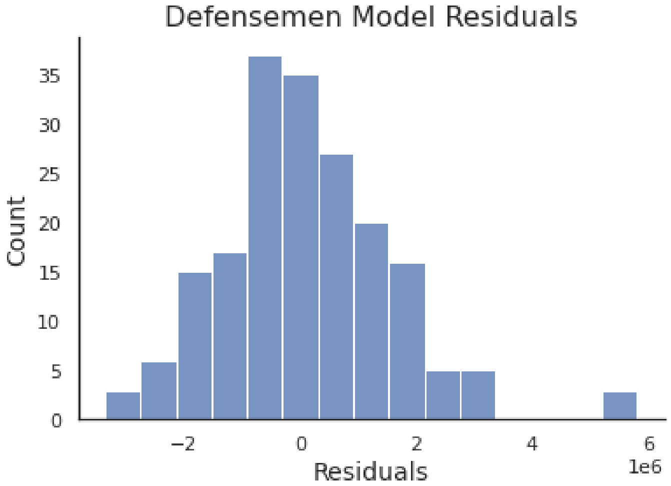


Figure 8. Histogram of Residuals for the Ridge Regression Defensemen Model.

Using the model to predict cap hit, we then compared the results to the actual cap hit and looked at the top 10 defensemen with the highest residual error (Fig. 9 and Fig. 10).

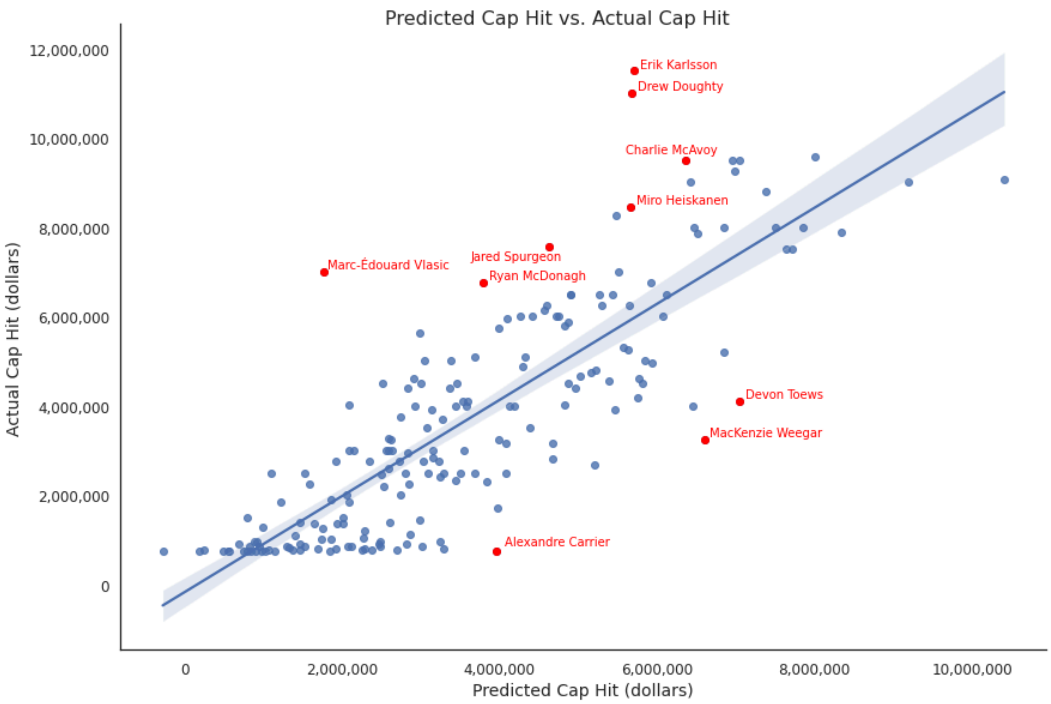


Figure 9. Predicted Cap Hit vs Actual Cap Hit for defensemen.

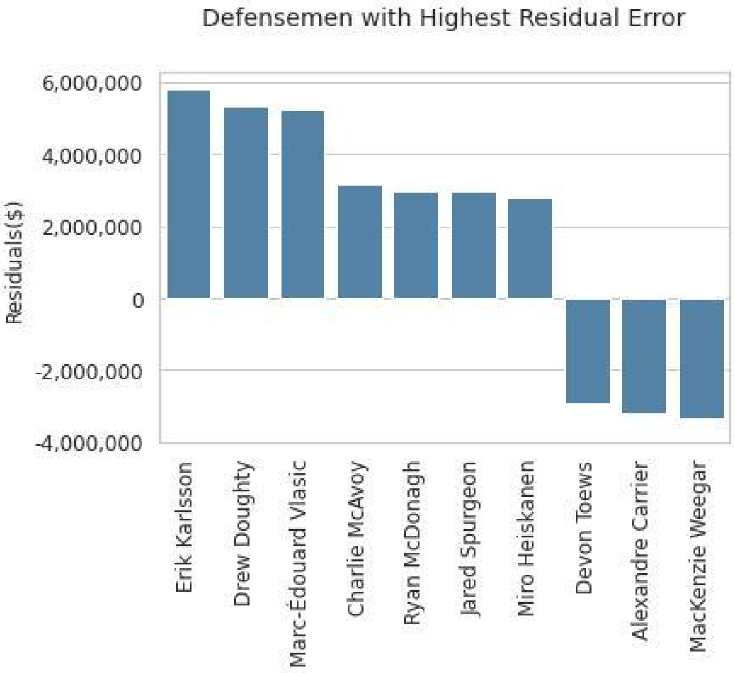


Figure 10. Defensemen with Highest Residual Error.

Highest Under-predicted Defensemen

Erik Karlsson has the highest residual error for defensemen with underpredicted cap hits. He is the highest paid defensemen in the league at $11.5M cap hit with an 8-year contract, so there is a lot of expectation for him to perform at a high level. He is a two-time winner of the Norris Trophy which is awarded to the league's top defensive player each season. However, he has suffered from several injuries and required several surgeries over the past eight years, so he hasn't been producing at the level he is expected to (Jones 2021). His predicted cap hit is $5.7M which is well under his actual cap hit. As such, he had a down year during the 2021-22 season. However, he also has leadership qualities as he was the previous captain for the Ottawa Senators. As mentioned previously, this is a player characteristic that on-ice performance stats do not take into account. So similar to the forward model, there are notable deficiencies in the defensemen model when it comes to players who have a long term contract, players who are dealing with injury issues, or players who have off-ice characteristics that provide leadership value to the team.

Drew Doughty has the second highest residual error for defensemen with underpredicted cap hits. He is the second highest paid defensemen in the league at $11M cap hit and also on an 8-year contract. Doughty was also suffering from multiple injuries during the 2021-22 season as he disclosed that he was "injured a lot of last year" (Hoven 2022). He had missed more than a month early in the season with a knee contusion, and also missed the final two months of the season due to a wrist injury (Dooley 2022). As such, he did not perform up to the expectations of a $11M cap hit player. However, like Karlsson, Doughty also has leadership qualities that aren't captured in the stats used in the model. He has won 2 Stanley Cups with the Kings and is an alternate captain for the team. Furthermore, he has also won a Norris Trophy, so his ceiling is high and his cap hit is likely to reflect this as well as his leadership and his previous cup wins. Again, reinforcing our findings about the limitations of the model since game statistics or demographic data will not capture these variables.

Highest Over-predicted Defensemen

Mackenzie Weegar has the highest residual error for defensemen with overpredicted cap hits. His actual cap hit is $3.25M and his predicted cap hit is approximately $6.6M resulting in a $3.35M residual error. In the 2021-22 season, he had his best season up until that point, scoring 44 points. The three seasons prior he had 36, 18, and 15 points, respectively. He was also at the end of his 3-year contract, so his actual cap hit was likely priced more in that range of a 15-20 point defenseman than a 30-40 point defensemen. He was traded to the Calgary Flames at the end of the 2021-22 season and signed an 8-year contract extension with them at a $6.25M cap hit. This value is much closer to our predicted value of $6.6M from the defensemen model. As such, we can see a practical example of how this model could potentially be used to determine a cap hit for teams and players who are negotiating a new contract deal or extension.

Notable Differences from Previous Research

Previous research favors random forest models for similar analysis which yields a slightly improved RMSE. In the case of forwards, we also observed slightly better RMSE and R-squared for the random forest models. However, as previously mentioned, we selected the linear regression model as our final model as the improvements in the random forest models were immaterial and we wanted to prioritize interpretability of the model. For our defensemen model, we found that a ridge regression model was the best fitting model, performing just as well or better than the random forest models. Prior research by Jensen and Warren (2022) and Nugent (2018) did not separate forwards and defensemen, so this is one difference in our modeling approach. Additionally, the analysis performed by Jensen and Warren (2022) included data from the 2010-11 season to the 2021-22 season for over 1,500 players. As such, the additional data could also explain why the random forest model performed better in their analysis.

Conclusions and Recommendations

Comparing the two models, the defensemen model performs better than the forward model; however, both models suffer from the same limitations which is the inability to predict the cap hit value that is attributable to injuries and additional off-ice characteristics. This makes sense as only on-ice game statistics and demographic data were used. However, using P/GP, A/GP, age, and xGF we were able to explain 63% of the variance observed in cap hit for forwards and 73% of the variance for defensemen using P/GP, Sh/GP, and CA. We believe these are promising results given that we only used data from a singular season. Therefore, we would seek to improve this model by including additional seasons in the datasets.

From our analysis we recommend prioritizing P/GP, A/GP, age, and xGF for evaluating forwards and P/GP, Sh/GP, and CA for evaluating defensemen. These models can be used as a baseline approach to determining a player's expected cap hit. However, there are quite a few limitations to this model as mentioned above. Although some of the residual errors in the model, due to these limitations, can actually be used as a negotiation opportunity. For example, if a player is having a good season and the model overpredicts their cap hit, the player's agent could use the overpredicted cap hit to negotiate a higher contract value when signing a new contract or contract extension. Similarly, an organization could use this to negotiate a lower contract value if the player is underperforming relative to their current cap hit. Additionally, isolating players that are outperforming their current cap hit could also provide potential trade target opportunities for an organization. As such, teams should seek to acquire these players in order to increase overall team performance at a discounted cap hit value.

In comparing the forward model with the defensemen model, the defensemen model performed better out of the two, according to our accuracy metrics. This is likely due in part to the fact that there was less variability in the performance metrics used in the defensemen dataset. It is also likely that the ridge regression model type performed the best on the defensemen dataset compared to the forward dataset as there was much more collinearity observed (Appendix A). An important limitation that is unique to ridge regression models and should be noted is the tradeoff between variance and bias. While the penalty term in the ridge regression equation reduces variance, it also increases the bias of the resulting coefficients (Ashok 2022). Striking the balance between reduced variance and increased bias is crucial because too much bias and too little variance can result in underfitting while too little bias and too much variance can result in overfitting (Paunikar 2018). As such, in future work, we'd use parameter tuning techniques to ensure the penalty parameter is optimized in our ridge regression equation rather than using the default penalty parameter value.

To improve these models, we would first recommend including additional seasons and taking a weighted average of the variables with the more recent seasons having a larger weight. This would help with predicting the cap hit for players who might be having an up or down season. As this data is publicly available on websites like capfriendly.com, additional resources wouldn't be required to do so. Additionally, we would also recommend including variables that are not directly related to on-ice performance, such as player health/injury status, leadership quality, and previous Stanley Cup championships. Such data is likely not publicly available or at least conveniently compiled, so additional resources and time would need to be devoted to collect and format the data in a way that is fit to be fed to the models.

We believe the model fails to adjust for long contract lengths. The most recent collective bargaining agreement between the NHL and the NHL Players' Association (NHLPA) permits contracts up to 8 years in length for restricted free agents and up to 7 years in length for unrestricted free agents (NHLPA 2020). We observed the highest residuals with players signed to maximum length contracts such as Eichel, Seguin, Toews, Karlsson, and Doughty, however, we still see issues with shorter contracts as well. In the case of Kadri, he was in the last year of his 6-year contract, so his 2021-22 cap hit might not have been an accurate representation of his current play. His 2021-22 cap hit is more representative of how he played 6 years ago when he first started playing in the league and did not produce as many points. Therefore, using these types of observations in the training data could present prediction issues. For the current models, we recommend to omit long term contracts greater than 6 years and evaluate these players separately. In the future, these prediction problems could potentially be addressed by reintroducing contracts greater than 6 years to the training set, but breaking them down into 2-3 year segments that are more representative of the statistics for players as time progresses.

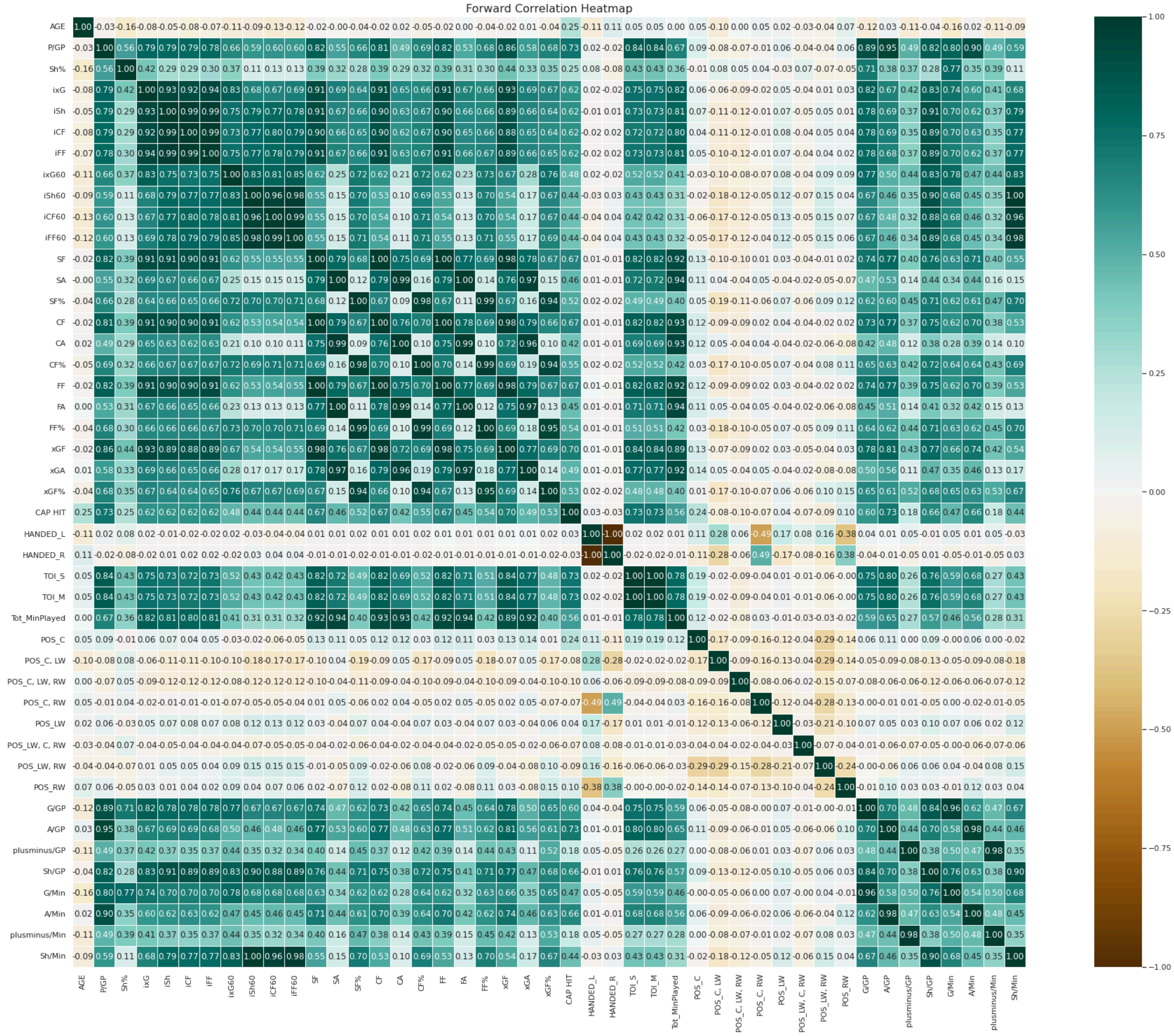
Another area to explore for further improvement would be to test a cubist model as this was shown to perform the best-fitting according to Jensen and Warren (2022). Due to time constraints and unfamiliarity with this type of modeling method, we did not try fitting this type of model on our datasets. Further research would be required in how to create such a model, but no additional resources as the same datasets would be used. Therefore, we also recommend exploring this type of modeling method to predict an NHL player's cap hit.

For this analysis, we excluded the goalie position because the position is significantly different from skaters which leads to different variables used to measure performance. Due to the nature of success in the position, there are few statistics available for goalies, and they are all centered around the same objective of shots saved. Additionally, there are typically only 2 goalies on each team and the majority of teams will play one of the goalies more than the other. As such, there is a much smaller sample size with several goalies having fewer games than the rest. Due to the limited number of goalies for each team, it's also important to consider a scenario of a goalie being injured. During those times a backup goalie may be called up from a lower league and then sent back down when the primary goalie has recovered. Therefore, we felt that using only the 2021-22 NHL season data would not produce a meaningful model. So, this is a model we would also be interested in exploring by using a weighted average of multiple seasons, and potentially different leagues, in order to include all positions on the team.

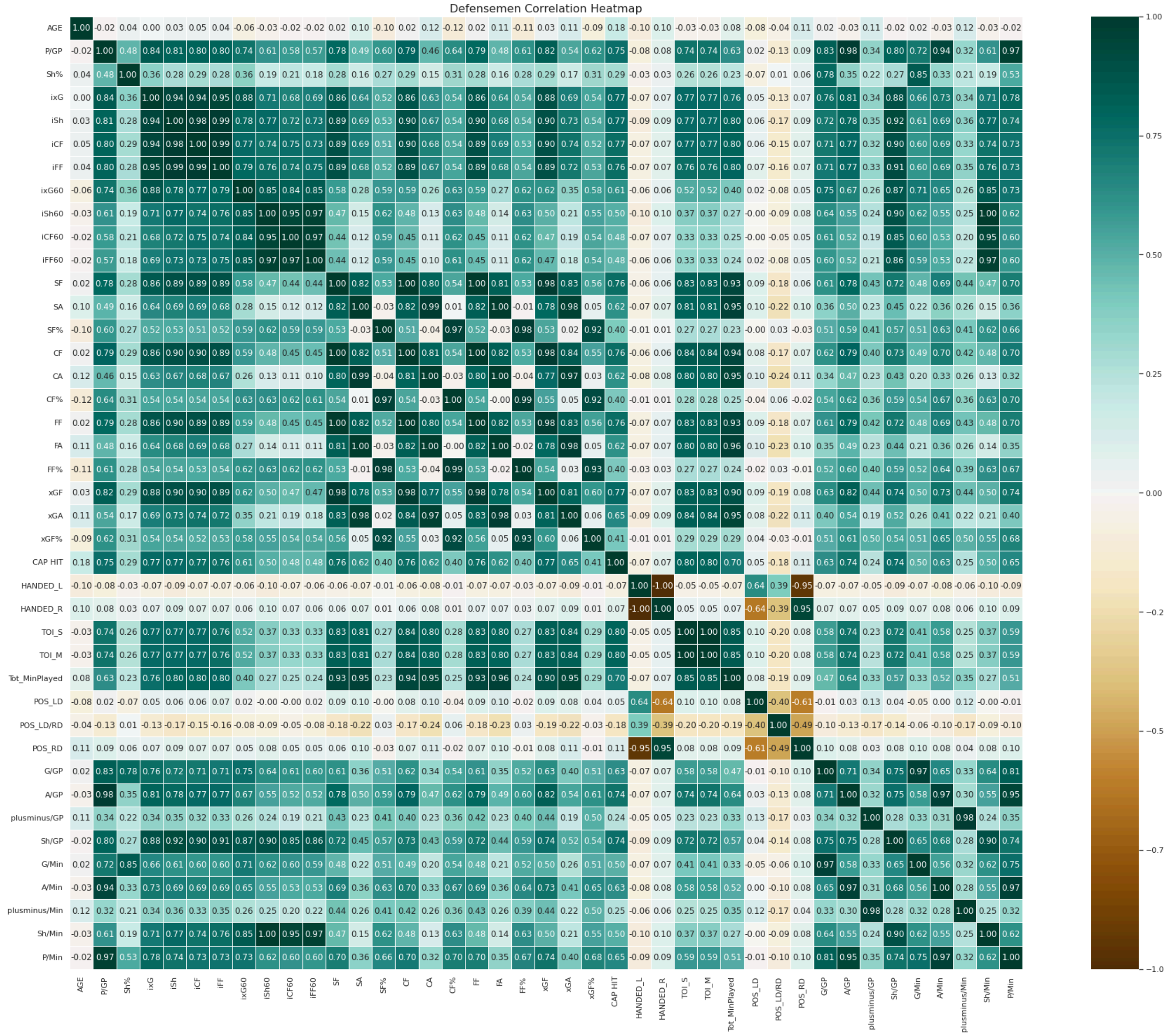
All in all, we recommend these models be used for initial considerations of cap hit allotment to players. As with all models, there are many limitations to be considered, thus the resulting cap hit predictions should be seen as a starting point for player valuation. Input from coaching staff, the general manager, players, front office staff and owners as well as negotiations will impact where the final number lands. With time, we hope to continually improve these models so they may require less human input after predictions and help organizations address the problem of player valuation through an efficient and data-driven approach.

Appendix A

The correlation matrix for the forward data set, as described in the “Methods” section, can be seen below:



Similarly, the correlation matrix for the defensemen data set can be seen below:


Defensemen Correlation Heatmap

	AGE	P/GP	Sh%	ixG	iSh	iCF	iFF	ixG60	iSh60	iCF60	iFF60	SF	SA	SF%	CF	CA	CF%	FF	FA	FF%	xGF	xGA	xGF%	CAP HIT	HANDED_L	HANDED_R	TOI_S	TOI_M	Tot_MinPlayed	POS_LD	POS_LD/RD	POS_RD	G/GP	A/GP	plusminus/GP	Sh/GP	G/Min	A/Min	plusminus/Min	Sh/Min	P/Min
AGE	1.00	-0.02	0.04	0.00	0.03	0.05	0.04	-0.06	-0.03	-0.02	-0.02	0.02	0.10	-0.10	0.02	0.12	-0.12	0.02	0.11	-0.11	0.03	0.11	-0.09	0.18	-0.10	0.10	-0.03	-0.03	0.08	-0.08	-0.04	0.11	0.02	-0.03	0.11	-0.02	0.02	-0.03	0.12	-0.03	-0.02
P/GP	-0.02	1.00	0.48	0.84	0.81	0.80	0.80	0.74	0.61	0.58	0.57	0.78	0.49	0.60	0.79	0.46	0.64	0.79	0.48	0.61	0.82	0.54	0.62	0.75	-0.08	0.08	0.74	0.74	0.63	0.02	-0.13	0.09	0.83	0.98	0.34	0.80	0.72	0.94	0.32	0.61	0.97
Sh%	0.04	0.48	1.00	0.36	0.28	0.29	0.28	0.36	0.19	0.21	0.18	0.28	0.16	0.27	0.29	0.15	0.31	0.28	0.16	0.28	0.29	0.17	0.31	0.29	-0.03	0.03	0.26	0.26	0.23	-0.07	0.01	0.06	0.78	0.35	0.22	0.27	0.85	0.33	0.21	0.19	0.53
ixG	0.00	0.84	0.36	1.00	0.94	0.94	0.95	0.88	0.71	0.68	0.69	0.86	0.64	0.52	0.86	0.63	0.54	0.86	0.64	0.54	0.88	0.69	0.54	0.77	-0.07	0.07	0.77	0.77	0.76	0.05	-0.13	0.07	0.76	0.81	0.34	0.88	0.66	0.73	0.34	0.71	0.78
iSh	0.03	0.81	0.28	0.94	1.00	0.98	0.99	0.78	0.77	0.72	0.73	0.89	0.69	0.53	0.90	0.67	0.54	0.90	0.68	0.54	0.90	0.73	0.54	0.77	-0.09	0.09	0.77	0.77	0.80	0.06	-0.17	0.09	0.72	0.78	0.35	0.92	0.61	0.69	0.36	0.77	0.74
iCF	0.05	0.80	0.29	0.94	0.98	1.00	0.99	0.77	0.74	0.75	0.73	0.89	0.69	0.51	0.90	0.68	0.54	0.89	0.69	0.53	0.90	0.74	0.52	0.77	-0.07	0.07	0.77	0.77	0.80	0.06	-0.15	0.07	0.71	0.77	0.32	0.90	0.60	0.69	0.33	0.74	0.73
iFF	0.04	0.80	0.28	0.95	0.99	0.99	1.00	0.79	0.76	0.74	0.75	0.89	0.68	0.52	0.89	0.67	0.54	0.89	0.68	0.54	0.89	0.72	0.53	0.76	-0.07	0.07	0.76	0.76	0.80	0.07	-0.16	0.07	0.71	0.77	0.33	0.91	0.60	0.69	0.35	0.76	0.73
ixG60	-0.06	0.74	0.36	0.88	0.78	0.77	0.79	1.00	0.85	0.84	0.85	0.58	0.28	0.59	0.59	0.26	0.63	0.59	0.27	0.62	0.62	0.35	0.58	0.61	-0.06	0.06	0.52	0.52	0.40	0.02	-0.08	0.05	0.75	0.67	0.26	0.87	0.71	0.65	0.26	0.85	0.73
iSh60	-0.03	0.61	0.19	0.71	0.77	0.74	0.76	0.85	1.00	0.95	0.97	0.47	0.15	0.62	0.48	0.13	0.63	0.48	0.14	0.63	0.50	0.21	0.55	0.50	-0.10	0.10	0.37	0.37	0.27	-0.00	-0.09	0.08	0.64	0.55	0.24	0.90	0.62	0.55	0.25	1.00	0.62
iCF60	-0.02	0.58	0.21	0.68	0.72	0.75	0.74	0.84	0.95	1.00	0.97	0.44	0.12	0.59	0.45	0.11	0.62	0.45	0.11	0.62	0.47	0.19	0.54	0.48	-0.07	0.07	0.33	0.33	0.25	-0.00	-0.05	0.05	0.61	0.52	0.19	0.85	0.60	0.53	0.20	0.95	0.60
iFF60	-0.02	0.57	0.18	0.69	0.73	0.73	0.75	0.85	0.97	0.97	1.00	0.44	0.12	0.59	0.45	0.10	0.61	0.45	0.11	0.62	0.47	0.18	0.54	0.48	-0.06	0.06	0.33	0.33	0.24	0.02	-0.08	0.05	0.60	0.52	0.21	0.86	0.59	0.53	0.22	0.97	0.60
SF	0.02	0.78	0.28	0.86	0.89	0.89	0.89	0.58	0.47	0.44	0.44	1.00	0.82	0.53	1.00	0.80	0.54	1.00	0.81	0.53	0.98	0.83	0.56	0.76	-0.06	0.06	0.83	0.83	0.93	0.09	-0.18	0.06	0.61	0.78	0.43	0.72	0.48	0.69	0.44	0.47	0.70
SA	0.10	0.49	0.16	0.64	0.69	0.69	0.68	0.28	0.15	0.12	0.12	0.82	1.00	-0.03	0.82	0.99	0.01	0.82	1.00	-0.01	0.78	0.98	0.05	0.62	-0.07	0.07	0.81	0.81	0.95	0.10	-0.22	0.10	0.36	0.50	0.23	0.45	0.22	0.36	0.26	0.15	0.36
SF%	-0.10	0.60	0.27	0.52	0.53	0.51	0.52	0.59	0.62	0.59	0.59	0.53	-0.03	1.00	0.51	-0.04	0.97	0.52	-0.03	0.98	0.53	0.02	0.92	0.40	-0.01	0.01	0.27	0.27	0.23	-0.00	0.03	-0.03	0.51	0.59	0.41	0.57	0.51	0.63	0.41	0.62	0.66
CF	0.02	0.79	0.29	0.86	0.90	0.90	0.89	0.59	0.48	0.45	0.45	1.00	0.82	0.51	1.00	0.81	0.54	1.00	0.82	0.53	0.98	0.84	0.55	0.76	-0.06	0.06	0.84	0.84	0.94	0.08	-0.17	0.07	0.62	0.79	0.40	0.73	0.49	0.70	0.42	0.48	0.70
CA	0.12	0.46	0.15	0.63	0.67	0.68	0.67	0.26	0.13	0.11	0.10	0.80	0.99	-0.04	0.81	1.00	-0.03	0.80	1.00	-0.04	0.77	0.97	0.03	0.62	-0.08	0.08	0.80	0.80	0.95	0.10	-0.24	0.11	0.34	0.47	0.23	0.43	0.20	0.33	0.26	0.13	0.32
CF%	-0.12	0.64	0.31	0.54	0.54	0.54	0.54	0.63	0.63	0.62	0.61	0.54	0.01	0.97	0.54	-0.03	1.00	0.54	-0.00	0.99	0.55	0.05	0.92	0.40	-0.01	0.01	0.28	0.28	0.25	-0.04	0.06	-0.02	0.54	0.62	0.36	0.59	0.54	0.67	0.36	0.63	0.70
FF	0.02	0.79	0.28	0.86	0.90	0.89	0.89	0.59	0.48	0.45	0.45	1.00	0.82	0.52	1.00	0.80	0.54	1.00	0.82	0.53	0.98	0.83	0.56	0.76	-0.07	0.07	0.83	0.83	0.93	0.09	-0.18	0.07	0.61	0.79	0.42	0.72	0.48	0.69	0.43	0.48	0.70
FA	0.11	0.48	0.16	0.64	0.68	0.69	0.68	0.27	0.14	0.11	0.11	0.81	1.00	-0.03	0.82	1.00	-0.00	0.82	1.00	-0.02	0.78	0.98	0.05	0.62	-0.07	0.07	0.80	0.80	0.96	0.10	-0.23	0.10	0.35	0.49	0.23	0.44	0.21	0.36	0.26	0.14	0.35
FF%	-0.11	0.61	0.28	0.54	0.54	0.53	0.54	0.62	0.63	0.62	0.62	0.53	-0.01	0.98	0.53	-0.04	0.99	0.53	-0.02	1.00	0.54	0.03	0.93	0.40	-0.03	0.03	0.27	0.27	0.24	-0.02	0.03	-0.01	0.52	0.60	0.40	0.59	0.52	0.64	0.39	0.63	0.67
xGF	0.03	0.82	0.29	0.88	0.90	0.90	0.89	0.62	0.50	0.47	0.47	0.98	0.78	0.53	0.98	0.77	0.55	0.98	0.78	0.54	1.00	0.81	0.60	0.77	-0.07	0.07	0.83	0.83	0.90	0.09	-0.19	0.08	0.63	0.82	0.44	0.74	0.50	0.73	0.44	0.50	0.74
xGA	0.11	0.54	0.17	0.69	0.73	0.74	0.72	0.35	0.21	0.19	0.18	0.83	0.98	0.02	0.84	0.97	0.05	0.83	0.98	0.03	0.81	1.00	0.06	0.65	-0.09	0.09	0.84	0.84	0.95	0.08	-0.22	0.11	0.40	0.54	0.19	0.52	0.26	0.41	0.22	0.21	0.40
xGF%	-0.09	0.62	0.31	0.54	0.54	0.52	0.53	0.58	0.55	0.54	0.54	0.56	0.05	0.92	0.55	0.03	0.92	0.56	0.05	0.93	0.60	0.06	1.00	0.41	-0.01	0.01	0.29	0.29	0.29	0.04	-0.03	-0.01	0.51	0.61	0.50	0.54	0.51	0.65	0.50	0.55	0.68
CAP HIT	0.18	0.75	0.29	0.77	0.77	0.77	0.76	0.61	0.50	0.48	0.48	0.76	0.62	0.40	0.76	0.62	0.40	0.76	0.62	0.40	0.77	0.65	0.41	1.00	-0.07	0.07	0.80	0.80	0.70	0.05	-0.18	0.11	0.63	0.74	0.24	0.74	0.50	0.63	0.25	0.50	0.65
HANDED_L	-0.10	-0.08	-0.03	-0.07	-0.09	-0.07	-0.07	-0.06	-0.10	-0.07	-0.06	-0.06	-0.07	-0.01	-0.06	-0.08	-0.01	-0.07	-0.07	-0.03	-0.07	-0.09	-0.01	-0.07	1.00	-1.00	-0.05	-0.05	-0.07	0.64	0.39	-0.95	-0.07	-0.07	-0.05	-0.09	-0.07	-0.08	-0.06	-0.10	-0.09
HANDED_R	0.10	0.08	0.03	0.07	0.09	0.07	0.07	0.06	0.10	0.07	0.06	0.06	0.07	0.01	0.06	0.08	0.01	0.07	0.07	0.03	0.07	0.09	0.01	0.07	-1.00	1.00	0.05	0.05	0.07	-0.64	-0.39	0.95	0.07	0.07	0.05	0.09	0.07	0.08	0.06	0.10	0.09
TOI_S	-0.03	0.74	0.26	0.77	0.77	0.77	0.76	0.52	0.37	0.33	0.33	0.83	0.81	0.27	0.84	0.80	0.28	0.83	0.80	0.27	0.83	0.84	0.29	0.80	-0.05	0.05	1.00	1.00	0.85	0.10	-0.20	0.08	0.58	0.74	0.23	0.72	0.41	0.58	0.25	0.37	0.59
TOI_M	-0.03	0.74	0.26	0.77	0.77	0.77	0.76	0.52	0.37	0.33	0.33	0.83	0.81	0.27	0.84	0.80	0.28	0.83	0.80	0.27	0.83	0.84	0.29	0.80	-0.05	0.05	1.00	1.00	0.85	0.10	-0.20	0.08	0.58	0.74	0.23	0.72	0.41	0.58	0.25	0.37	0.59
Tot_MinPlayed	0.08	0.63	0.23	0.76	0.80	0.80	0.80	0.40	0.27	0.25	0.24	0.93	0.95	0.23	0.94	0.95	0.25	0.93	0.96	0.24	0.90	0.95	0.29	0.70	-0.07	0.07	0.85	0.85	1.00	0.08	-0.19	0.09	0.47	0.64	0.33	0.57	0.33	0.52	0.35	0.27	0.51
POS_LD	-0.08	0.02	-0.07	0.05	0.06	0.06	0.07	0.02	-0.00	-0.00	0.02	0.09	0.10	-0.00	0.08	0.10	-0.04	0.09	0.10	-0.02	0.09	0.08	0.04	0.05	0.64	-0.64	0.10	0.10	0.08	1.00	-0.40	-0.61	-0.01	0.03	0.13	0.04	-0.05	0.00	0.12	-0.00	-0.01
POS_LD/RD	-0.04	-0.13	0.01	-0.13	-0.17	-0.15	-0.16	-0.08	-0.09	-0.05	-0.08	-0.18	-0.22	0.03	-0.17	-0.24	0.06	-0.18	-0.23	0.03	-0.19	-0.22	-0.03	-0.18	0.39	-0.39	-0.20	-0.20	-0.19	-0.40	1.00	-0.49	-0.10	-0.13	-0.17	-0.14	-0.06	-0.10	-0.17	-0.09	-0.10
POS_RD	0.11	0.09	0.06	0.07	0.09	0.07	0.07	0.05	0.08	0.05	0.05	0.06	0.10	-0.03	0.07	0.11	-0.02	0.07	0.10	-0.01	0.08	0.11	-0.01	0.11	-0.95	0.95	0.08	0.08	0.09	-0.61	-0.49	1.00	0.10	0.08	0.03	0.08	0.10	0.08	0.04	0.08	0.10
G/GP	0.02	0.83	0.78	0.76	0.72	0.71	0.71	0.75	0.64	0.61	0.60	0.61	0.36	0.51	0.62	0.34	0.54	0.61	0.35	0.52	0.63	0.40	0.51	0.63	-0.07	0.07	0.58	0.58	0.47	-0.01	-0.10	0.10	1.00	0.71	0.34	0.75	0.97	0.65	0.33	0.64	0.81
A/GP	-0.03	0.98	0.35	0.81	0.78	0.77	0.77	0.67	0.55	0.52	0.52	0.78	0.50	0.59	0.79	0.47	0.62	0.79	0.49	0.60	0.82	0.54	0.61	0.74	-0.07	0.07	0.74	0.74	0.64	0.03	-0.13	0.08	0.71	1.00	0.32	0.75	0.58	0.97	0.30	0.55	0.95
plusminus/GP	0.11	0.34	0.22	0.34	0.35	0.32	0.33	0.26	0.24	0.19	0.21	0.43	0.23	0.41	0.40	0.23	0.36	0.42	0.23	0.40	0.44	0.19	0.50	0.24	-0.05	0.05	0.23	0.23	0.33	0.13	-0.17	0.03	0.34	0.32	1.00	0.28	0.33	0.31	0.98	0.24	0.35
Sh/GP	-0.02	0.80	0.27	0.88	0.92	0.90	0.91	0.87	0.90	0.85	0.86	0.72	0.45	0.57	0.73	0.43	0.59	0.72	0.44	0.59	0.74	0.52	0.54	0.74	-0.09	0.09	0.72	0.72	0.57	0.04	-0.14	0.08	0.75	0.75	0.28	1.00	0.65	0.68	0.28	0.90	0.74
G/Min	0.02	0.72	0.85	0.66	0.61	0.60	0.60	0.71	0.62	0.60	0.59	0.48	0.22	0.51	0.49	0.20	0.54	0.48	0.21	0.52	0.50	0.26	0.51	0.50	-0.07	0.07	0.41	0.41	0.33	-0.05	-0.06	0.10	0.97	0.58	0.33	0.65	1.00	0.56	0.32	0.62	0.75
A/Min	-0.03	0.94	0.33	0.73	0.69	0.69	0.69	0.65	0.55	0.53	0.53	0.69	0.36	0.63	0.70	0.33	0.67	0.69	0.36	0.64	0.73	0.41	0.65	0.63	-0.08	0.08	0.58	0.58	0.52	0.00	-0.10	0.08	0.65	0.97	0.31	0.68	0.56	1.00	0.28	0.55	0.97
plusminus/Min	0.12	0.32	0.21	0.34	0.36	0.33	0.35	0.26	0.25	0.20	0.22	0.44	0.26	0.41	0.42	0.26	0.36	0.43	0.26	0.39	0.44	0.22	0.50	0.25	-0.06	0.06	0.25	0.25	0.35	0.12	-0.17	0.04	0.33	0.30	0.98	0.28	0.32	0.28	1.00	0.25	0.32
Sh/Min	-0.03	0.61	0.19	0.71	0.77	0.74	0.76	0.85	1.00	0.95	0.97	0.47	0.15	0.62	0.48	0.13	0.63	0.48	0.14	0.63	0.50	0.21	0.55	0.50	-0.10	0.10	0.37	0.37	0.27	-0.00	-0.09	0.08	0.64	0.55	0.24	0.90	0.62	0.55	0.25	1.00	0.62
P/Min	-0.02	0.97	0.53	0.78	0.74	0.73	0.73	0.73	0.62	0.60	0.60	0.70	0.36	0.66	0.70	0.32	0.70	0.70	0.35	0.67	0.74	0.40	0.68	0.65	-0.09	0.09	0.59	0.59	0.51	-0.01	-0.10	0.10	0.81	0.95	0.35	0.74	0.75	0.97	0.32	0.62	1.00

References

Ashok, Prasanth. 2022. “What is Ridge Regression?” Great Learning, November 16, 2022. https://www.mygreatlearning.com/blog/what-is-ridge-regression/.

Beheshti, Nima. 2022. “Random Forest Regression: A basic explanation and use case in 7 minutes.” Towards Data Science, March 2, 2022. https://towardsdatascience.com/random-forest-regression-5f605132d19d.

Brownlee, Jason. 2020. “Ordinal and One-Hot Encodings for Categorical Data.” Machine Learning Mastery, June 12, 2020. https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/.

Cohen, Joshua. 2023. “Chicago Blackhawks Star Jonathan Toews Is Stepping Away from Hockey, Says He's Dealing with Long Covid.” Forbes. February 22. https://www.forbes.com/sites/joshuacohen/2023/02/19/chicago-blackhawks-star-jonathan-toews-is-stepping-away-from-hockey-says-hes-dealing-with-long-covid/?sh=32f33c1c5c5e.

Dooley, Zach. 2022. “21/22 Seasons in Review – Drew Doughty.” *LA Kings Insider*. June 21. https://lakingsinsider.com/2022/06/22/21-22-seasons-in-review-drew-doughty/.

Généreux, Louis-Charles and Allen Xu. 2021. “DETERMINING NHL PLAYER SALARIES AND ASSESSING GENERAL MANAGER EFFECTIVENESS WITH MACHINE LEARNING.” Northwestern University: MsiA Student Research, June 4, 2021. https://sites.northwestern.edu/msia/2021/06/04/determining-nhl-player-salaries-and-assessing-general-manager-effectiveness-with-machine-learning/.

Goldman, Shayna. 2022. “Analytics in the NHL: Where It's at and Where It's Going Next.” *The Athletic*. June 24. https://theathletic.com/3381342/2022/06/24/nhl-analytics-stats/.

Hoven, John. 2022. “Doughty: Hurt or Healed, He's Still Offering Advice.” *Mayor's Manor.com*. September 29. https://mayorsmanor.com/2022/09/doughty-already-itching-to-get-going-it-hurt-not-playing/.

Jensen, Haily, and Eric Warren. 2022. “NHL Salary Cap Project.” Carnegie Mellon University Statistics & Data Science, July 29, 2022. https://www.stat.cmu.edu/cmsac/sure/2022/showcase/nhl_salary.html.

Jones, C.G. 2021. “Sharks Need Karlsson to Build upon Surgery-Free Offseason.” *The Hockey Writers*. August 2. https://thehockeywriters.com/sharks-karlsson-2021-surgery-free-offseason-opportunities/.

Jones, Wayne. n.d. “What is the lowest salary in the NHL? (and who is making it).” Hockey Answered. Accessed February 16, 2023. https://hockeyanswered.com/what-is-the-lowest-salary-in-the-nhl/.

Murphy, Bryan. 2022. “NHL entry-level contract, explained: How much can rookies make on their first NHL deal?” The Sporting News, July 7, 2022. https://www.sportingnews.com/us/nhl/news/nhl-entry-level-contract-explained-rookies-deal/tb61ploxqpandyrvktnv33ti#:~:text=NHL%20entry%2Dlevel%20contracts%20and%20salaries&text=Any%20player%20younger%20than%2025,is%20set%20at%20%24925%2C000%20annually.

National Hockey League Players’ Association (NHLPA). 2020. “Collective Bargaining Agreement (CBA)” NHLPA: Collective Bargaining Agreement. Accessed March 7, 2022. https://www.nhlpa.com/the-pa/cba.

Nugent, Cam. 2018. “NHL Salary Data Prediction: Cleaning and Modeling.” Kaggle, January 26, 2018. https://www.kaggle.com/code/camnugent/nhl-salary-data-prediction-cleaning-and-modeling/notebook.

Paunikar, Ankita. 2018. “Intuition behind Bias-Variance trade-off, Lasso and Ridge Regression.” Data Science Central, April 1, 2018. https://www.datasciencecentral.com/intuition-behind-bias-variance-trade-off-lasso-and-ridge/.

Ravanshad, Abolfazl. 2018. “Gradient Boosting vs Random Forest.” Medium, April 27, 2018. https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80.

Schram, Carol. 2022. “Report: NHL Salary Cap Projected To Make $4 Million Jump For 2024-25 Season.” Forbes, September 27, 2022. https://www.forbes.com/sites/carolschram/2022/09/27/report--nhl-salary-cap-projected-to-make-4-million-jump-for-2024-25-season/?sh=59d356066895.

Severini, Thomas A. 2020. *Analytic Methods in Sports: Using Mathematics and Statistics to Understand Data from Baseball, Football, Basketball, and Other Sports*. Boca Raton, FL: CRC Press.

Webster, Danny. 2021. “Jack Eichel Undergoes Successful Neck Surgery.” *Knights On Ice*. November 12. https://www.knightsonice.com/2021/11/12/22779007/vegas-golden-knights-center-jack-eichel-successful-neck-surgery-artificial-disk-replacement.