# Predicting & Evaluating NHL Player Cap Hits

BrainStation, April 2022
Brian Johns
the.brian.johns@gmail.com

## PROBLEM STATEMENT

---

Every NHL team has the goal to win the Stanley Cup. To do so, the league enforces financial restrictions that every team has to abide by. For the 2021-22 season, the main restrictions are:

1. Roster Limit - **20-23 Active Players**
2. Team Salary Cap - **$81.5 million total** for all active players.
3. Max Salary - **$16.3 million**
4. 'Entry-Level' Contracts - Maximum entry-level contract is **$925,000** for the first 3 years of their career.
5. Minimum Salary - **$700,000**.
6. Cap Hit Calculation - The **average salary across the full duration of a player's contract**.

Concurrently, the use of advanced analytics has increased in popularity over the last decade. The amount of data available for each player is immense and any team who can effectively use this data may have an advantage in player acquisition and performance.

**So how can we predict and evaluate the Cap Hit of NHL players using their basic and advanced statistics?**

Answering this will allow us to:

1. Identify statistics that players are currently being rewarded for.
2. Identify statistics that may be undervalued
3. Identify players that are under/overvalued based on their performance profile.

## BACKGROUND

---

The hockey analytics community has strived to properly measure an individual player's contribution to the success of a team. Some measures have included:

Measuring the quality of possession for teams and players.
Measuring a player's contributions into a single number
Visualizing the impact of a player on the ice

The NHL has lagged behind other professional sports in measuring and collecting data in this area. However, information has been gathered for long enough now that there is enough data to collect and analyze in hopes of validating these measures along with performance.

Machine learning and data science techniques can enable the hockey analytics community to use these statistics to better evaluate and analyze player performance. By analyzing how a player's cap hit and salary is determined by their statistics, this project will hopefully add strategies for teams to identify, acquire and retain talent that will lead to on ice success.

# DATA SOURCE

There were two primary sources of data for this project:

1. Evolving Hockey

Evolving Hockey stores a myriad of hockey statistics. Of these, I selected 8 tables in 4 categories:

Goals Above Replacement Statistics
Expected Goals Above Replacement Statistics
Box Score Statistics
On Ice Statistics

Each of these categories had one table where statistics were measured as totals and one table that measured each statistics as a rating: per 60 minutes of Ice Time. Definitions of each of these statistics can be found in the attached Glossary.

Subscribers to the website are able to download the data in .csv format. I downloaded the last 10 seasons of data was downloaded on March 25th, 2022.

2. CapFriendly

Webscraping was used to acquire financial information for all players from CapFriendly. 10 seasons of data was acquired by scraping across multiple pages (50 players per web page). The data I used for this project was scraped on March 25th, 2022.

In order to merge this data together, player names were formatted to allow the financial information of a player to align with their player performance statistics. Goalies were not included in this project.

# PRE-PROCESSING

In preparation for modeling, I performed exploratory data analysis to evaluate the distribution of the data and to gain initial insights about what factors may correlate with the Cap Hit of a player.

Most of the performance statistics had a normal distribution. However, 24 statistics had either exponential or bimodal distributions. These statistics were log-transformed in order to distribute the data normally in preparation for modeling.

The majority of statistics showed a positive correlation with Cap Hit. The strongest positive correlations were with Goals For and Expected Goals For.

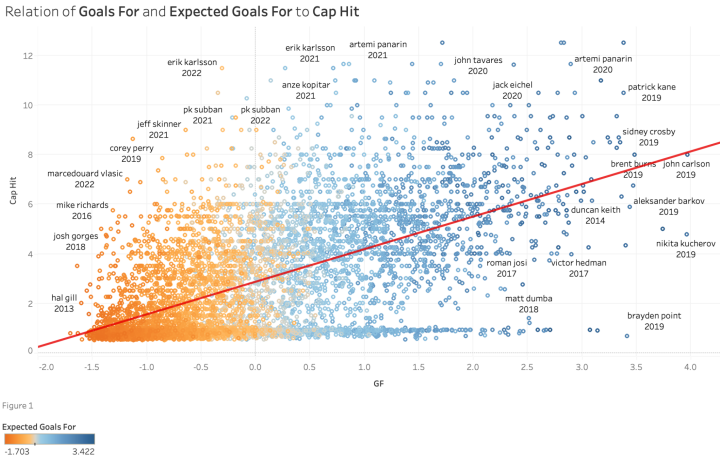


Figure 1

A few statistics did show a negative correlation with Cap Hit, such as advanced statistics measuring the defensive contributions of a player.

There is a distinct collection of players who make less than $1 million that have a large range of performance levels. These are players that have ‘entry-level’ contracts as young players, or are playing at the veteran minimum.

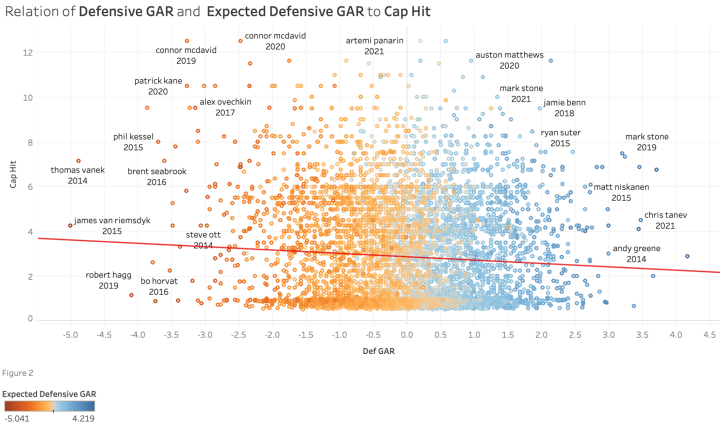



Figure 2

# REGRESSION MODEL SUMMARY

---

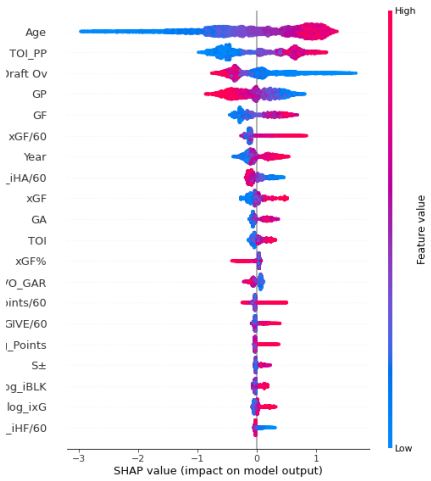
I used a series of regression models in order to predict the Cap Hit of NHL players through their performance statistics. These included:

- Linear Regression
- Linear Regression with Ridge and Lasso Regularization
- XGBoost Regressor
- Random Forest Regressor

Multicollinearity was a concern in performing these models due to the number of statistics available. Through feature selection and Principal Component Analysis, I reduced the dimensionality of the data and re-ran each of the regression models above.

The model that provided the best results was an XGBoost Regressor model performed on the untransformed dataset. The scores for this model were:

R-Squared: 0.6790
Mean Absolute Error: 0.9398
Root Mean Squared Error: 1.2905.




Even though the XGBoost model had the highest predictive score, all of the regression models shared very similar insights as to what features were the most important in predicting a player’s Cap Hit. The most important features in predicting a player’s Cap Hit included:

1. Age
2. Time On Ice for the Powerplay
3. Draft Pick - Overall
4. Goals For
5. Expected Goals For

This highlights that factors outside of performance influence Cap Hit more than performance factors.

In evaluating the XGBoost model, two themes emerged where the model was not able to predict the Cap Hit of the players:

1.Players with ‘mega’ contracts (over $10 million)

2.Players on ‘minimum’ or ‘entry-level’ contracts that out-performed their contract.

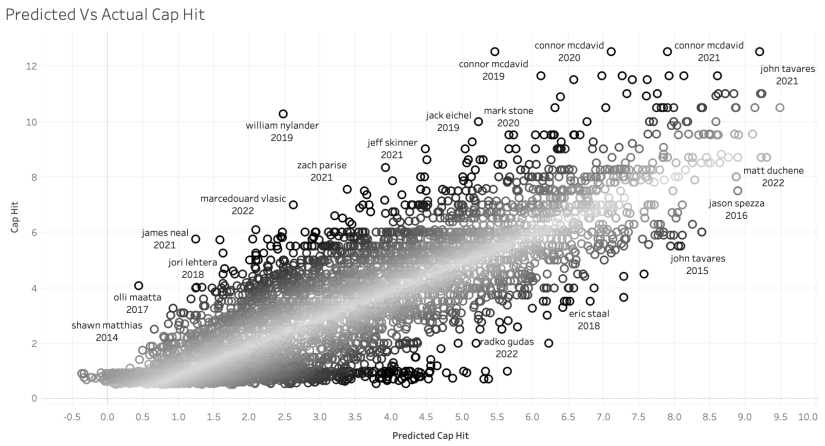



Figure 3 - Shading Shows the Difference Between the Predicted and Actual Cap Hit Values

# CLUSTERING SUMMARY

K-Means Clustering was used to group NHL players together by their statistics. Clustering all of the players in the dataset yielded low silhouette scores (0.19 for 2 clusters, 0.11 for 4 clusters) without any actionable information gained.

K-Means Clustering was run again, but this time separating players by position: Forwards and Defencemen. This returned a similarly low silhouette score for both positions, but with clusters that contained more actionable information. This is how the positions were clustered:

**Forwards**:
#1 - Star Forwards
#2 - Offensive Producer, Defensive Liability
#3 - Primary Defensive Forwards
#4 - Secondary Defensive Forwards
#5 - Fringe NHL Forwards
#6 - Enforcers

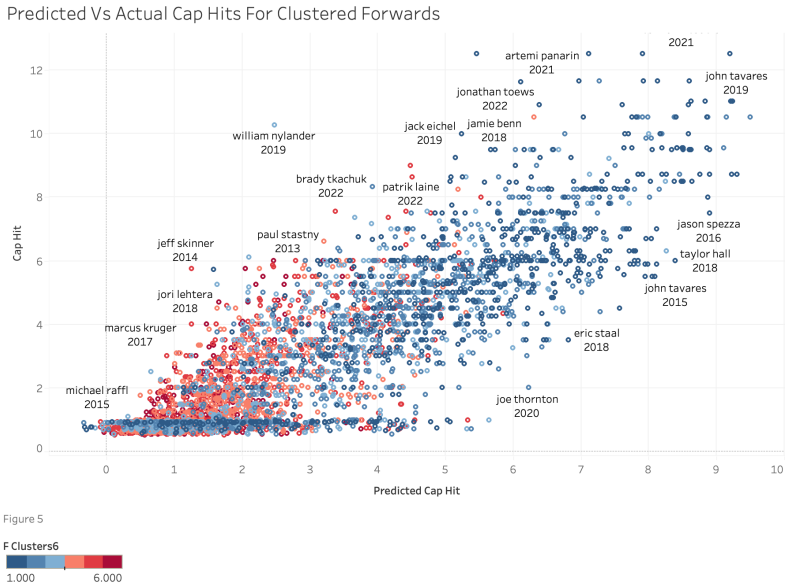



Figure 5

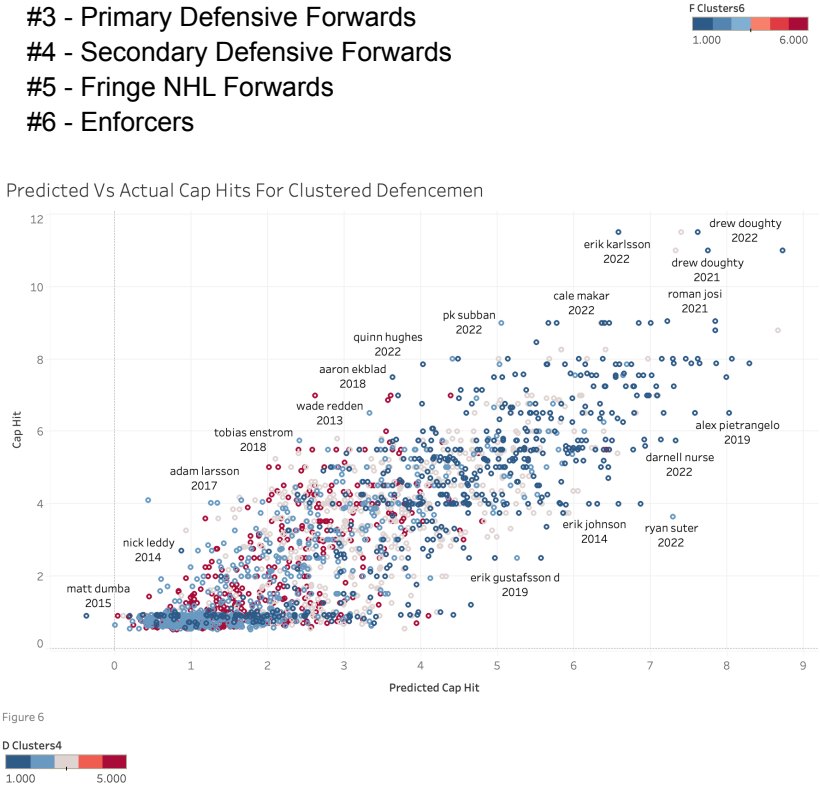



Figure 6

**Defencemen:**
#1: Star Defencemen
#2: Offensive Producer in Limited Time
#3: Defensive Defencemen
#4: Fringe NHL Defencemen

Players with the highest cap hits in the ‘Fringe’ clusters were players on the back end of long-term contracts that were performing poorly relative to their contact. Conversely, there was a very high number of players in the ‘Star’ clusters who were on ‘entry-level’ or minimum contracts.

## FINDINGS

1. Non-performance related features, such as Overall Draft Pick and Age, are the most important in determining a player's cap hit. Goals For and Expected Goals For were the most important performance factors.

2. Advanced statistics, such as Goals Above Replacement, were not important when measuring the Cap Hit of a player, particularly defensive metrics. Great offensive players, even if they are poor defensively, are still rewarded significantly.

3. There's significant value in finding good players that make less than $1 million. Emphasizing drafting and developing good young players can enable a team to spend money elsewhere. Less obviously, identifying aging veterans, with past success but at low cost, that can still perform at a high level is potentially an effective team-building strategy.

4. The most overvalued players were players that were on the end of very long-term contracts and no longer capable of performing to the level of the contract.

5. Potentially, based on the models, it could be that **NO** player may be worth a contract worth more than $10 million based on how much measurable differentiation there is between players at that level.

However, the predictions and clustering in these models were relatively weak with many gaps to fill. Even the best model cannot explain 33% of the variance, and some incalculable value (leadership, marketability, quality of agent) could be a more decisive factor.

## SUMMARY

This project hints at some meaningful insights, especially emphasizing talent evaluation for players that are available for less than $1 million (rookies and veterans). However, finding that no player is worth more than $9 million would require more definitive support before being considered a legitimate strategy.

The findings of this project suggest two competing team-building strategies:

Pay over $10 million for superstars, then invest heavily into developing young players and veterans on minimum salaries.

**OR**

Acquire several $4-6 million players that can outperform their salaries, even at the cost of trading a superstar.

These potential next steps may provide better team-building strategies going forward:
1. Re-focus this project on a small subset of features, ie. GAR Ratings
2. Run a similar project with Draft Pick as the target variable to identify draft specific team-building strategies
3. Develop PALM, or Price Above League Minimum, to convert a player's stats into a dollar value, then compare that to their actual salary.