Using Machine Learning Algorithms to Predict Salaries in the National Hockey League (NHL)

By Matthew Gilbody

04/21/2023

California State University, Dominguez Hills

Department of Computer Science

Committee Chair: Dr. Jianchao “Jack” Han

Committee Members: Dr. Bin Tang and Dr. Alexander Chen

Abstract

The world of analytics has grown over the last few years. Most companies are making decisions based off of data driven insights and the major sports leagues are no different. Ever since the release of *Moneyball*, the idea of using statistics to find value in players has become increasingly popular. Teams across all the major sports leagues are investing heavily into analytics and using the insights that they gain to make decisions, such as, strategy, marketing, salary, trades, etc. This paper aims to take data from the current NHL season and use various machine learning algorithms, such as, random forest, lasso regression, and multiple linear regression in order to predict player salary based on player statistics. We will look at how each algorithm works, how well they performed, and what factors go into making their predictions.

I. Introduction

Performance is a big factor when it comes to financial compensation. No matter the profession, performance is often analyzed and can lead to increased compensation or being terminated, and nowhere is this more true than the world of sports. Players are signed to term contracts and the value of those contracts is highly dependent on what kind of value the player brings to the franchise. Player salary has become an intense topic. Players across the various major sports leagues feel that their value has increased and have turned to tactics such as requesting trades, holding out, and refusing to engage in team activities in order to increase pressure on the team to give them the financial compensation that they feel they deserve. On the other side of the coin, teams often have to navigate hard salary caps, making it tricky to sign top players, while still fielding competitive teams. This paper is going to focus on the National Hockey League (NHL). The goal of this paper is to find out what goes into determining a player's salary. In the upcoming sections, I will discuss the data that is used and how the data has been transformed and broken down. I will also discuss the various algorithms that were used to obtain the predictions, the results, and the conclusions that I drew from the project.

II. The Problem and Background

The NHL is one of four major sports leagues in the United States and pulls in billions of dollars each year. While still not as popular as football, baseball, or basketball, the league has boosted its popularity and according to Forbes, "Hockey team owners are scoring big, with the average NHL team value now $1.03 billion, topping $1 billion for the first time and 19% more than a year ago" ("NHL Team Values 2022", 2022). Hockey has always been a violent sport, but

with the influx of highly skilled players and lower levels of fighting, hockey has seen an increase in popularity. According to a Frontiers article by Fortney et. Al, “NHL fighting rates have diminished during the past two decades, while fan attendance has increased. A significant negative correlation exists between fan attendance and fights per game, casting doubt on fighting’s entertainment value” (Fortney et. Al). With this influx of talent, comes an influx in player salaries. According to a hockeyskillstraining.com article, “In 1990, Wayne Gretzky was the highest paid player in the NHL at $3 million per season” (“NHL Salaries Over the Years”) and according to the data set used in this project, the average salary is around $3 million. Wayne Gretzky is considered to be the greatest player to ever step on the ice, and even though $3 million was considered great money in 1990, when it is adjusted for inflation to the modern day, it’s only around $7 million. Some players are currently making $12.5 million and that number is only going to increase as time goes on.

The biggest problems that teams have to face is constructing a competitive roster while staying within the constraints of the salary cap. This can be hard to reconcile as investing a lot of money into a few players can have lasting impacts on the long term success of a franchise. Allocating large sums of money to a few players limits the flexibility a team has, meaning they’ll have to sign more average performing players. As previously mentioned, I am going to explore what goes into determining a player’s salary. The goal of this paper is to determine which factors are most important in determining a player’s salary, which players are underpaid/overpaid, and if there are other factors outside of player performance that go into determining a player’s value.

III. Data

The data for this project was collected from hockey-reference.com and includes players' basic statistics, advanced statistics and their respective cap hit from the current 2022-2023 season. I chose cap hit versus their actual salary, because cap hit is the annual average value (AAV) of their contract. Using their salary isn't as efficient, because the way some contracts are set up, a player's actual salary can be different each year depending on the structure of the contract. I broke up the data into 3 separate parts: defense, centers and wingers. I did this because each of these positions contributes differently and I wanted to see how differently each algorithm would determine the variables that are most important for that position. Each of these positions could be broken down even more. I could separate based upon the lines they play on, how much power play time the players get, handedness they are, where they were drafted, etc. but I chose to not break it down that far. Separating them into their broad positional categories is sufficient for this study. Players who have played less than 10 games were excluded from the project as that is not enough data to draw any meaningful conclusions. Goalies were not included in this project as well. In the sections that follow, I will mainly show visualizations from the NHL Centers data. This is to reduce redundancy and to make this paper more concise. I will share the results for all positions in the results section.

The data was cleaned using excel. I was able to connect to the web and load the data into an excel worksheet. From there, I was able to merge queries, rename the columns, delete certain columns, and separate the various positions. Once I was satisfied with the data, I converted the file from an excel workbook to a CSV (comma separated file). Another method of doing this is by using the pandas library. I mainly used pandas to read in the CSV files, get the basic data from the file, show a correlation matrix, and format the results of the predictions. Pandas is a

very powerful data manipulation tool, but I chose to work in excel to clean the data. This allowed me to focus on implementing the algorithms on the data.

Figure 1 (NHL Centers Data Summary)

	Age	GP	G	PLMI	CAPHIT
count	222.000000	222.000000	222.000000	222.000000	2.220000e+02
mean	27.279279	68.621622	16.000000	-1.108108	3.117166e+06
std	4.189345	15.586602	11.558625	14.193142	2.770609e+06
min	19.000000	11.000000	0.000000	-38.000000	4.500000e+05
25%	24.000000	60.250000	7.000000	-9.000000	8.941670e+05
50%	27.000000	74.000000	13.500000	-1.000000	1.762500e+06
75%	30.000000	81.000000	22.000000	8.000000	5.000000e+06
max	38.000000	84.000000	64.000000	42.000000	1.250000e+07

Figure 1 represents a small summary of the "NHL Centers" data table. Centers have the most responsibility on the ice as they are tasked with face offs, providing offense, while also helping on defense. Figure 1 gives a breakdown of 5 categories from the data on centers. There are 222 centers that were a part of this study. I would expect faceoff win percentage, their scoring ability, their plus/minus, and some defensive stats to play prominent roles in determining their value. We can see that the highest paid center is $12.5 million, while the lowest paid center is $450,000 and the average sits at just above $3 million.

Figure 2 (NHL Wingers Data Summary)

	Age	GP	G	PLMI	CAPHIT
count	196.000000	196.000000	196.000000	196.000000	1.960000e+02
mean	27.448980	64.923469	15.326531	-1.239796	3.271559e+06
std	3.867345	18.454358	11.153728	13.182687	2.630222e+06
min	20.000000	12.000000	0.000000	-33.000000	7.500000e+05
25%	25.000000	56.000000	7.000000	-9.000000	9.250000e+05
50%	27.000000	72.000000	13.000000	-3.000000	2.500000e+06
75%	30.000000	80.000000	21.000000	7.000000	5.135417e+06
max	38.000000	84.000000	61.000000	41.000000	1.164286e+07

Figure 2 represents a small summary of the “NHL Wingers” data table. Wingers play alongside centers and are mainly responsible for providing offense. They do have some defensive responsibility, but wingers are more expected to provide scoring. We have slightly less players in this data pool than we do centers, but we see that there are similar averages across the board. For wingers, I would expect goals scored, assists, points, plus/minus, and other offensive minded stats will be more prominent in determining their value.

Figure 3 (NHL Defense Data Summary)

	Age	GP	PTS	PLMI	CAPHIT
count	228.000000	228.000000	228.000000	228.000000	2.280000e+02
mean	27.754386	63.403509	22.399123	2.298246	3.030652e+06
std	4.025385	19.548750	17.487904	15.225141	2.536367e+06
min	20.000000	10.000000	1.000000	-41.000000	7.333330e+05
25%	24.000000	51.000000	10.000000	-7.250000	8.599998e+05
50%	28.000000	70.500000	18.000000	3.000000	2.347075e+06
75%	31.000000	79.000000	31.000000	12.000000	4.500000e+06
max	39.000000	85.000000	101.000000	49.000000	1.150000e+07

Figure 3 represents a small summary of the “NHL Defense" data table. Defensemen focus on keeping the puck out of their own net more than they do about putting the puck in the net of

their opponent. They are the last line of defense before it gets to the goalie. Higher paid defensemen bring a higher grade of offense than most defensemen in the league, but it isn't their primary responsibility. We see similar averages in age and cap hit for all positions, but we won't see the same when it comes to scoring. For defensemen, I would expect to see plus/minus, total time on ice, age, and other defensive stats to be more prominent in determining their value.

IV. Algorithms and Methods

Random Forest

The first method that I used was the random forest algorithm. Random Forest is as exactly as the name implies. It's a collection of decision trees where independent variables are chosen at random and split until a final node is reached. Once each tree has made its prediction, it will combine the results from all the trees and output a final prediction. Random Forest is considered to be one of the most effective algorithms when it comes to classification and regression problems. The reason being, features are randomly selected, regardless of which features are deemed the most important, making each tree independent of one another. Li writes, "RF will take random selection for features (predictors in the research, like age, goal.....) rather than always using all features to train the decision trees" (Li, pg. 10). In this case, the final prediction is a result of the average taken from all the individual predictions from each tree. If we were doing classification, then the final prediction would be the result of a majority vote. As previously mentioned, there are 3 different categories of NHL players. Each category was run through the algorithm.

Figure 4 (Decision Tree from our Random Forest)

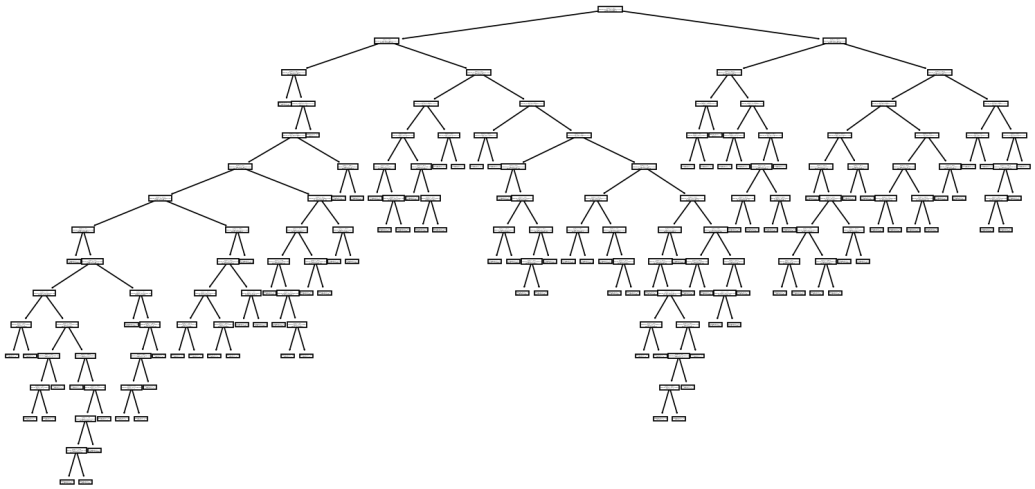
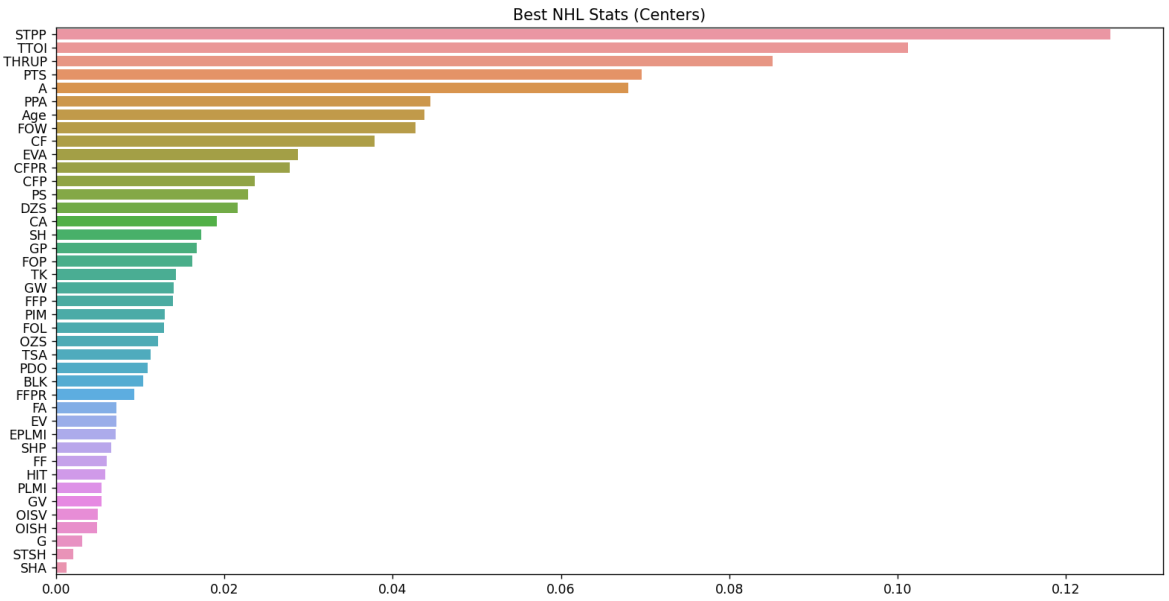


Figure 4 is an example of a decision tree from our Random Forest. Unfortunately, the image was too large and I couldn't see what each node was split at, but it offers a glimpse into the structure of the tree and what happens during the process. Each tree begins with a randomly chosen condition and bakes a binary split at each node until an ending condition is reached. Once each tree outputs its prediction, they can all be averaged out to form the final prediction. For this project, I created 100 trees for all centers, wingers, and defensemen. I used 80% of the data as

the training data and the remainder was used as the test data. This was done for centers, wingers, and defensemen.

Figure 5 (Feature Selection for Centers for our Random Forest)



The above graph depicts the most important features ranked from most important to least important for centers in the NHL. There are similar figures for wingers and defensemen. The top 3 features that the model selected for centers are power play goals (STPP), total time on ice (TTOI), and thru percentage (THRUP), while the bottom 3 features are goals (G), short-handed goals (STSH), and short-handed assists (SHA). While face-off win (FOW) looked to be among the better features, I would have expected it to rank in the top 3. Centers have a large responsibility with face-offs. Face-offs really impact the game in many ways, so a player with a high face-off win percentage proves himself to be highly effective. The model was run once before feature selection was performed and then once after. Once the features were selected and ranked, the bottom two features were dropped to see if that improved our initial prediction.

Multiple Linear Regression

Multiple Linear Regression (MLR) is a machine learning algorithm that takes more than one independent variable to make a prediction on the dependent variable. In this case, our dependent variable is the player's cap hit and the independent variables are the quantifiable stats, such as the amount of goals they scored. In their paper, Tonack used a MLR model for evaluating centers. Tonack states, "The multiple linear regression model solves the question of an NHL center's value by predicting salary based on various on-ice performance metrics by taking the respective coefficient values of each independent variable in the regression and combining them to determine the deserved salary based on the player's statistics" (Tonack). Tonack describes doing an experiment where 17 out of a possible 500 variables are used to predict a player's salary. The variables that were chosen need to have a decent correlation between them and the player's salary, because in order for MLR to work, variables need to be chosen based upon how they correlate to the dependent variable. The equation looks like this with 'Y' representing our dependent variable, 'B0' being our intercept, and 'B1/Bn' represent the coefficient for each independent variable, 'X1/Xn' represent our independent variables, and 'e' represents the margin of error:

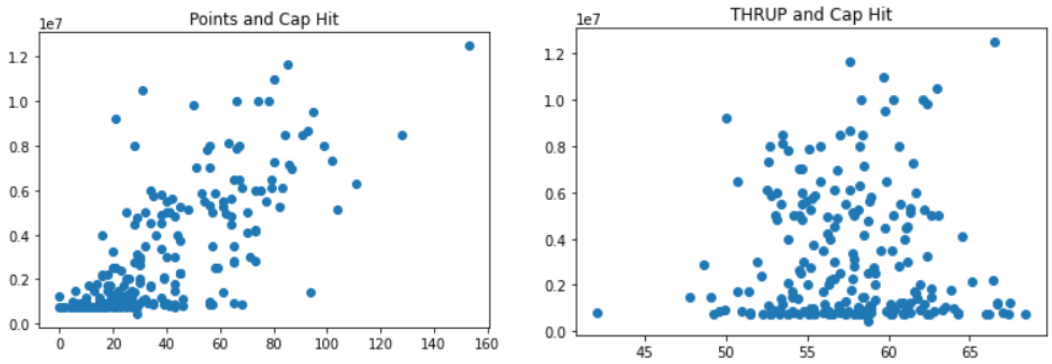
$$Y = B0 + B1X1 + BnXn + e$$

Figure 6 (Correlation Map for Centers)

	Age	GP	G	A	PTS	PLMI	PIM	PS	EV	STPP
Age	1.000000	0.088038	0.069710	0.071823	0.073987	0.033233	0.051328	0.068181	0.014404	0.137425
GP	0.088038	1.000000	0.567018	0.579177	0.598760	0.113784	0.288527	0.515304	0.586903	0.361020
G	0.069710	0.567018	1.000000	0.834177	0.942402	0.232818	0.087300	0.957289	0.934347	0.833833
A	0.071823	0.579177	0.834177	1.000000	0.970596	0.253488	0.100895	0.917308	0.725150	0.774358
PTS	0.073987	0.598760	0.942402	0.970596	1.000000	0.255360	0.099297	0.974179	0.847622	0.833594
PLMI	0.033233	0.113784	0.232818	0.253488	0.255360	1.000000	-0.157389	0.402802	0.262763	0.102188
PIM	0.051328	0.288527	0.087300	0.100895	0.099297	-0.157389	1.000000	0.042473	0.093332	0.050562
PS	0.068181	0.515304	0.957289	0.917308	0.974179	0.402802	0.042473	1.000000	0.874949	0.827301
EV	0.014404	0.586903	0.934347	0.725150	0.847622	0.262763	0.093332	0.874949	1.000000	0.595108
STPP	0.137425	0.361020	0.833833	0.774358	0.833594	0.102188	0.050562	0.827301	0.595108	1.000000

Figure 6 represents a partial snippet of a correlation map between all the different variables. The closer the score is to 1, the more the two variables show correlation between them. Unfortunately, I couldn't show the entire correlation map due to size constraints, but for purposes of this paper, showing a snippet will be sufficient. I ran the algorithm once without removing any of the variables. Once I produced this correlation map, I was able to see which variables did not show a correlation with the cap hit. I removed the variables that showed a negative correlation score and ones that were less than 0.2. By doing this, I was able to use only the variables that showed some amount of correlation, which would hopefully improve the results of the prediction.

Figure 7 (Scatterplots of highly correlated variable and variable with no correlation)



The above two scatter plots show 2 different variables, points(PTS) and thru percentage (THRUP) and how they correlate to cap hit for centers. As mentioned above, the algorithm was run with all the variables included and once again with the variables that showed a correlation score of less than 0.2. The MLR model showed that mostly offensive stats for centers showed the most correlation with a player's cap hit. Face-off stats showed some correlation, but it was not as high as I would have expected. Later on, we analyze the results of this algorithm.

Lasso Regression

Lasso Regression works similarly to linear regression, but aims to improve it. According to Glen, "Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean" (Glen). In our MLR model, the coefficient values aren't normalized, whereas in Lasso regression, they are. This means that the coefficient values are moved towards a value for 0. Lasso uses alpha, which serves to "punish" the model. If a coefficient reaches a value of 0, then we can remove that coefficient from the equation and get a more optimal prediction. For the Lasso model, I started by assigning an alpha score of 0.1. Once I got the first prediction, I seeked to get a more optimal alpha score to see if that would improve upon the prediction. The process is very similar to the MLR model, the only difference is, the process of scaling down the coefficients and getting an alpha score that is most optimal for our data set.

V. Results

A. NHL Centers

Random Forest Results

```
First Prediction vs New Prediction
       Salary  First Prediction  Difference 1  New Prediction  Difference 2
210  10500000        5341268.10   -5158731.90      5443929.77   -5056070.23
152   2500000        3062616.27     562616.27      2924750.02     424750.02
135    925000        2091700.42    1166700.42      2008390.44    1083390.44
122   7800000        5156288.99   -2643711.01      5074284.50   -2725715.50
162    775000        1519287.53     744287.53      1608141.70     833141.70
-----------------------------------
First R^2 value: 0.7089967111164871
New R^2 value: 0.703197375501976
```

When the test is first run, we see the r-squared value is around 0.71 and doesn't really change much when we take out the variables that don't have much effect on predicting the cap hit. Because the r-squared value is high, we can be confident that the random forest model was effective in making salary predictions. In this iteration, face-off stats and some offensive categories were near the top, meaning that the model selects more offensive categories for centers as opposed to a mixture of both offensive and defensive stats. We can also conclude that the random forest model is not affected by eliminating less correlated stats.

Multiple Linear Regression Results

```
      Salary  First Prediction  Difference 1  Second Prediction  Difference 2
177  4000000      2.693520e+06  1.306480e+06       2.781398e+06  1.218602e+06
34   1125000      4.276592e+05  6.973408e+05       5.898233e+05  5.351767e+05
89   4500000      3.609401e+06  8.905993e+05       3.690375e+06  8.096254e+05
116   863333      4.882561e+05  3.750769e+05       3.920999e+05  4.712331e+05
67    750000      1.980548e+06 -1.230548e+06       2.218674e+06 -1.468674e+06
First R^2 = 0.31743718564886425
New R^2 = 0.3861949620335262
```

As we can see from the results, the MLR model was not very effective in predicting player salary. One key difference from the random forest model is the fact that the model did improve once we removed the stats that had little correlation with the dependent variable. The MLR model focused heavily on offensive stats, and felt that face-off percentage wasn't as big of a factor as stats such as goals scored, assists and points. Because of this model's lack of flexibility, it contributed to less effective results.

Lasso Regression Results

```
['Age', 'GP', 'A', 'PTS', 'PLMI', 'STPP', 'PPA', 'S', 'BLK', 'FOW', 'OISV ', 'TK', 'GV', 'EPLMI', 'THRUP']
      Salary  First Prediction  ...  New Prediction  Second Difference
130   766667      2.324867e+05  ...    4.107614e+05       3.559056e+05
203  2750000      2.424278e+06  ...    1.644988e+06       1.105012e+06
170   925000      3.472901e+06  ...    2.675665e+06      -1.750665e+06
66    800000      3.166553e+06  ...    2.905836e+06      -2.105836e+06
181  6125000      4.975696e+06  ...    5.673656e+06       4.513437e+05

[5 rows x 5 columns]
First R^2 = 0.5084007628960814
Second R^2 = 0.6253258674409218
```

The Lasso model showed a big improvement once we removed the columns that did not have an effect on our final prediction. The first iteration gave us an r-squared of around and jumped up to .62 once we found our optimal alpha. This model favored offensive stats, but we also see that face-off wins(FOW), was a contributing factor to the model. FOW is a stat that I would expect to play a prominent role in all models as it is an important part of a center's job on the ice. I feel confident that the Lasso model is a good predictor for a player's salary. While it was not as effective as the random forest model, the Lasso model showed a big improvement over the MLR model, which can be attributed to the "punishment" factor that the Lasso model provides.

B. NHL Wingers

Random Forest Results

```
First Prediction vs New Prediction
      Salary  First Prediction  Difference 1  New Prediction  Difference 2
68   2375000        3570975.00    1195975.00      3364392.08     989392.08
2     785000        1500933.74     715933.74      1578733.74     793733.74
136  4750000        3125067.94   -1624932.06      3193264.37   -1556735.63
70   3150000        5981894.64    2831894.64      5903238.80    2753238.80
156  2500000        2495667.90      -4332.10      2430717.08     -69282.92
First R^2 value: 0.6172528297445279
New R^2 value: 0.6291112906020584
```

For our wing players, we get a fairly decent r-squared value, meaning that we can say that the random forest algorithm was fairly effective in predicting player salary. As expected, the model felt that offensive stats were the most effective in predicting the player's salary. Wing players are more offensive minded players, and stats like total shots, shot percentage, points, and goals were big factors in producing these results. Another interesting observation from these results is that the model did improve after we removed the least important features. Although it didn't improve by much, it is still an improvement and runs contrary to the random forest algorithm run for centers.

Multiple Linear Regression Results

	Salary	First Prediction	Difference 1	Second Prediction	Difference 2
127	835833	6.952434e+05	1.405896e+05	4.587068e+05	3.771262e+05
11	1000000	1.492187e+06	-4.921866e+05	2.196933e+06	-1.196933e+06
110	5250000	2.878219e+06	2.371781e+06	2.384124e+06	2.865876e+06
124	3000000	6.659274e+06	-3.659274e+06	6.138679e+06	-3.138679e+06
18	5400000	3.402453e+06	1.997547e+06	3.421048e+06	1.978952e+06

First R^2 = 0.33677533565418094

New R^2 = 0.3386093850688354

As we can see from the results, MLR was again, not effective in determining player salary. A common theme between centers and wingers is the low correlation between plus/minus (PLMI), which is a little surprising for centers since PLMI is a stat that measures both your offense and defensive effectiveness. For wingers, defense isn't their first priority. Most wingers in the NHL are there to score goals and provide effective offensive play. I would expect their PLMI to be lower than both defense and centers. One surprising observation is to see that there was virtually no improvement in the model once the lowest correlated variables were removed. I would assume that the model would show some sort of improvement since this model relies heavily on correlated stats.

Lasso Regression Results

```
['Age', 'GP', 'A', 'PLMI', 'STSH', 'EVA', 'SP', 'CF', 'FA', 'OZS ', 'GV', 'TSA']
      Salary  First Prediction  ...  New Prediction  Second Difference
83   2625000      7.505020e+06  ...    6.894327e+06      -4.269327e+06
12   1200000      1.230420e+06  ...    1.050258e+06       1.497421e+05
33   4000000      2.421590e+06  ...    3.664933e+06       3.350669e+05
113   750000      7.647115e+05  ...    1.155211e+06      -4.052106e+05
171  3250000      1.256779e+06  ...    1.749042e+06       1.500958e+06

[5 rows x 5 columns]
First R^2 = -0.3970528203842314
Second R^2 = -0.037675854928953756
```

The Lasso model for wingers was not effective at all as we got a negative r-squared value. This is surprising, considering that the MLR model produced an r-squared value of 0.33. I would expect to see an improvement, just like we saw with the center's data since the Lasso model has the "punishment" feature to it. The negative r-squared means that we can't draw any meaningful conclusions and can confidently say that the Lasso model is highly ineffective in producing a prediction for wingers in the NHL.

C. NHL Defense

Random Forest Results

```
First Prediction vs New Prediction:
      Salary  First Prediction  Difference 1  New Prediction  Difference 2
17    900000        2373313.34    1473313.34      2675571.67    1775571.67
140  2000000        3694215.50    1694215.50      3889295.34    1889295.34
226  1137500         772141.65    -365358.35       776625.00    -360875.00
148   925000        1072866.66     147866.66      1023374.99      98374.99
168  6750000        5138113.58   -1611886.42      5145493.40   -1604506.60
R^2 value: 0.545711130201925
New R^2 value: 0.5351142125691171
```

For our defensive players, our model didn't perform as well as it did for centers and wingers. While 0.54 isn't a terrible r-squared score, it isn't enough to say that the random forest

model was very effective in predicting salary. Unlike the wingers, the model performed less effectively when the least important features were removed from consideration.

Multiple Linear Regression Results

```
       Salary  First Prediction  Difference 1  Second Prediction  Difference 2
39     750000      8.113659e+05 -6.136586e+04       6.264417e+05  1.235583e+05
169   6150000      3.478960e+06  2.671040e+06       3.732333e+06  2.417667e+06
93    7875000      5.626141e+06  2.248859e+06       6.026962e+06  1.848038e+06
62    3250000      3.100824e+06  1.491756e+05       2.445002e+06  8.049984e+05
198   5300000      5.320395e+06 -2.039461e+04       4.554373e+06  7.456274e+05
First R^2 = 0.575889410751836
New R^2 = 0.5882138077613444
```

The MLR has a slight edge over the random forest model as the r-squared value was a little closer to 0.6. While the results are still somewhat mediocre, we can say that we are somewhat confident that the MLR is effective in predicting player salaries for defensive players. It is surprising that the MLR model performed better than the random forest model. The random forest model proved to be effective in predicting the salaries for both wingers and centers, but was less effective with defense players. Given the fact that random forest is able to pull features randomly from the data set and each tree is independent, it would be safe to assume that the random forest would have outperformed the MLR model.

Lasso Regression Results

```
['Age', 'GP', 'A', 'PLMI', 'PIM', 'PS', 'STPP', 'STSH', 'GW', 'SHA', 'S', 'SP', 'TTOI', 'BLK', 'HIT', 'FFPR', 'OISH ', 'OISV ', 'OZS ', 'TK', 'GV', 'EPLMI']
      Salary  First Prediction  ...  New Prediction  Second Difference
139  9500000      4.238510e+06  ...    4.489925e+06       5.010075e+06
200  2750000      1.609576e+06  ...    1.953997e+06       7.960033e+05
122  1400000      1.740916e+06  ...    2.231264e+06      -8.312638e+05
185   850000      1.189863e+06  ...    1.943429e+06      -1.093429e+06
141  6750000      5.803816e+06  ...    5.180130e+06       1.569870e+06

[5 rows x 5 columns]
First R^2 = 0.5927328184670142
Second R^2 = 0.613231921042215
```

The Lasso model performed slightly better than the MLR model, but all 3 models didn't separate themselves from the rest. The Lasso model was the best model for prediction for

defensive players, while random forest was the best for centers and wingers. An interesting observation from this result is the amount of features that were included in the new prediction. The wingers model only featured 12 stats, the centers model only featured 15, but the defense model used 22 stats. The other models seemed to favor offensive stats in the final predictions, even for defensive players, but the Lasso model used stats such as PLMI, PIM, TTOI, BLK, and HIT, which match up to what I think are some of the most important stats for defensemen.

VI. Conclusion

This project consisted of using 3 different machine learning algorithms to predict salary for NHL players. The Random Forest model seemed to have the best results for centers and wingers, while the Lasso algorithm was the best predictor for defensemen. Random Forest is one of the most effective machine learning algorithms and I would conclude that the Random Forest model was the most effective for this project. The reason being is that features are chosen at random for each tree and is more representative of the data. The MLR was the least effective model overall, which is surprising, because the features that didn't show much correlation to our dependent variable were removed, but the model never improved. This model depends on correlation between the independent variables and dependent variables in order to be effective, and the fact that it performed poorly is surprising. Finally, the Lasso model seemed to work effectively for both defense and centers, but provided no insights for wingers. This was also surprising as the MLR model provided us with a positive r-squared value and the Lasso model seeks to improve upon the linear regression model.

A limitation of this study was my decision to leave players with entry level contracts. Other studies argued that having them in would reduce the accuracy of the model because their

salary isn't dependent on their performance. The problem with this is that some players are near the end of their entry level contracts and could be entering free agency where they can sign a higher value deal, meaning that their performance does affect their salary.

This project's goal was to see how stats were used to determine what a player should be paid. While the models showed effectiveness, I would argue that stats alone cannot be the only factor that goes into determining a player's salary. Other factors are fan attendance, a player's jersey sales, their leadership, feeling amongst teammates and coaches, and what the player means to the city and organization that he plays for. Jersey sales and ticket sales are quantifiable stats and could improve upon determining a player's salary, but what the player means to the city, leadership traits and respect from players and coaches can be a determining factor as well, and these stats cannot be quantified. What fans see is the production, or lack thereof, on the ice, but the intangible factors play a bigger role than we might think. There is no doubt that there is a correlation between a player's production and their salary, but their salary can't be solely based on how they play.

Works Cited

Fortney, Thomas, et al. "National Hockey League Fights per Game and Viewership Trends: 2000-2020." *Frontiers*, 30 Jun. 2022, www.frontiersin.org/articles/10.3389/fspor.2022.890429/full#:~:text=Conclusions%3A%20NHL%20fighting%20rates%20have,doubt%20on%20fighting's%20entertainment%20value. Accessed 20 Apr. 2023.

Glen, Stephanie. "Lasso Regression: Simple Definition" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/lasso-regression/

Li, Chenyao, et al. *Machine Learning Modeling to Evaluate the Value of Football Players*. 2022. University College London, Final Project.

Morthi, Aparna. "How Lasso Regression Works in Machine Learning." *Dataaspirant*, 26 Nov. 2020, dataaspirant.com/lasso-regression/#t-1606404715787. Accessed 19 Apr. 2023.

Ozanian, Mike, and Justin Teitelbaum. "NHL Team Values 2022." *Forbes*, 22 Dec. 2022, www.forbes.com/sites/mikeozanian/2022/12/14/nhl-team-values-2022-new-york-rangers-on-top-at-22-billion/?sh=eb74d967deb1. Accessed 18 Apr. 2023.

Tonack, Austin. *Determining an NHL Center'S Value: Salary Prediction Based on Performance Data*. 2018. Project.

hockey-reference.com

"NHL Salaries Over the Years." *HockeySkillsTraining*, www.hockeyskillstraining.com/nhl-salaries-over-the-years/. Accessed 19 Apr. 2023.

Glossary of Terms

A	assists - first or second person to pass to the eventual goal scorer

Age	age - how old the person is
BLK	blocks - how many shots a person blocked
CA	corsi
CAPHIT	cap hit - a player's aav of their contract
CF	corsi at even strength (shots+blocks+misses)
CFP	corsi for percentage
CFPR	relative corsi for percentage
DZS	defensive zone start percentage
EPLMI	expected plus minus
EV	even strength goals - how many goals a player has when it is 5 on 5
EVA	even strength assists - how many assists a player has when it is 5 on 5
FA	fenwick against (shots + misses)
FF	fenwick for (shots+misses)
FFP	fenwick for percentage

FFPR	relative fenwick for percentage
G	goals - total number of goals a player has scored
GP	games played - total number of games a player has played
GV	giveaways
GW	game winning goals - how many goals a player has that resulted in winning the game
HIT	hit - how many hits or checks a player has
OISH	team on ice shooting percentage when player is on the ice
OISV	team on ice save percentage when player is on the ice
OZS	offensive zone start percentage
PDO	shooting percentage plus save percentage
PIM	penalty in minutes - total number of minutes a player spends in the penalty box
Player	player - name of the

	player
PLMI	plus minus
Pos	position - what position a player plays
PPA	power play assists - how many assists a player has on the power play
PS	point share
PTS	points - goals plus assists
S	shots - how many shots a player has
SHA	short handed assists - how many assists a player has when they are short handed
SP	shot percentage - total goals divided by total shots taken
STPP	special teams power play - how many goals a player has while their team is on the power play
STSH	special teams shots - how many shots a player has on the power play
THRUP	thru percentage - number of shots that record as a shot on

	goal
TK	takeaways
Tm	team - what team a player plays for
TSA	total shots attempted
TTOI	total time on ice - total minutes a player has spent on the ice for the season