USING MACHINE LEARNING ALGORITHMS TO PREDICT SALARIES IN THE NATIONAL HOCKEY LEAGUE

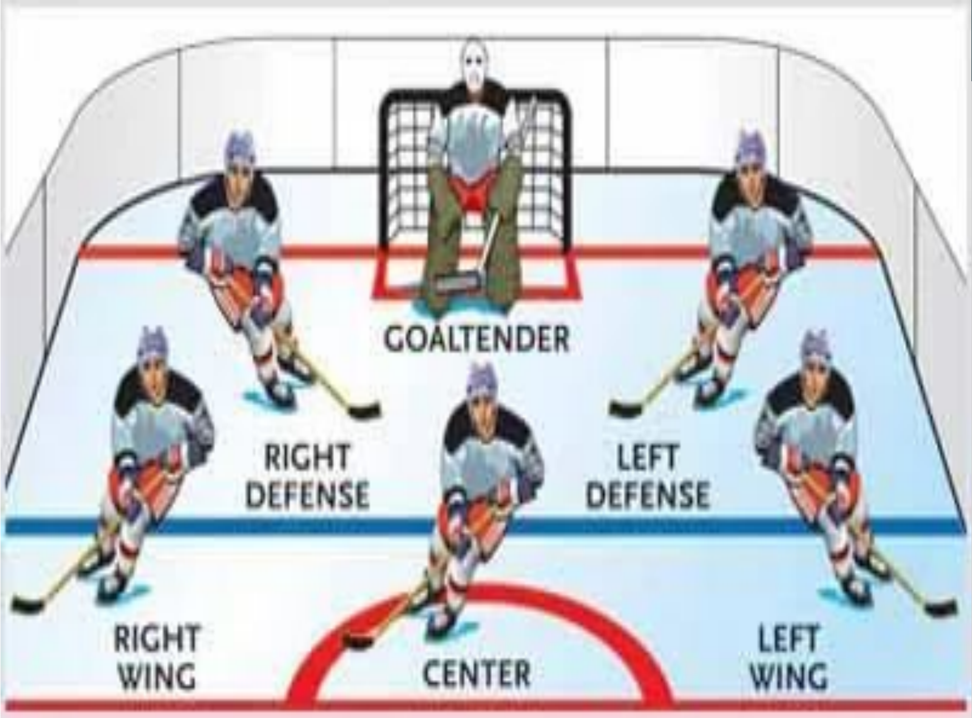
By Matthew Gilbody

Project Rundown

- Introduction
  - Background
  - Motivations
- The data
  - How it was split
  - How it was structured
  - Why it was structured that way
- Algorithms
- Conclusions
- Code demonstration
- Q&A

Introduction

- Hockey is played with a total of 6 skaters on the ice
  - 1 goalie, 1 centerman, 2 wingers and 2 defensemen
- Salaries range from $750k - $12.5 million
- Hockey generates about $5 billion annually
- Grown in popularity since the reduction in fighting and increase in skill
- With the increase in skill comes greater spending on players



Motivations

- Data analytics has become a major part of the sports industry
- Used for business strategies, finding value in players, and on-field/on-ice strategy
- Popularized by Bill James and later on with the movie *Moneyball* (OBP)
- Effects long term success of the organization

The Data

- Split into 3 categories
  - Defense
  - Wingers
  - Centers
- Each position contributes differently and has various responsibilities
- Better to compare players to other players who play the same position
- Data was gathered from:
  - https://www.hockey-reference.com/leagues/NHL_2023_skaters.html
  - https://www.hockey-reference.com/leagues/NHL_2023_skaters-advanced.html
  - https://www.hockey-reference.com/friv/current_nhl_salaries.cgi
- Data was cleaned and prepared in Excel
- Dependent variable will be Cap Hit which represents the AAV (average annual value of the contract)

Data (Center Summary)

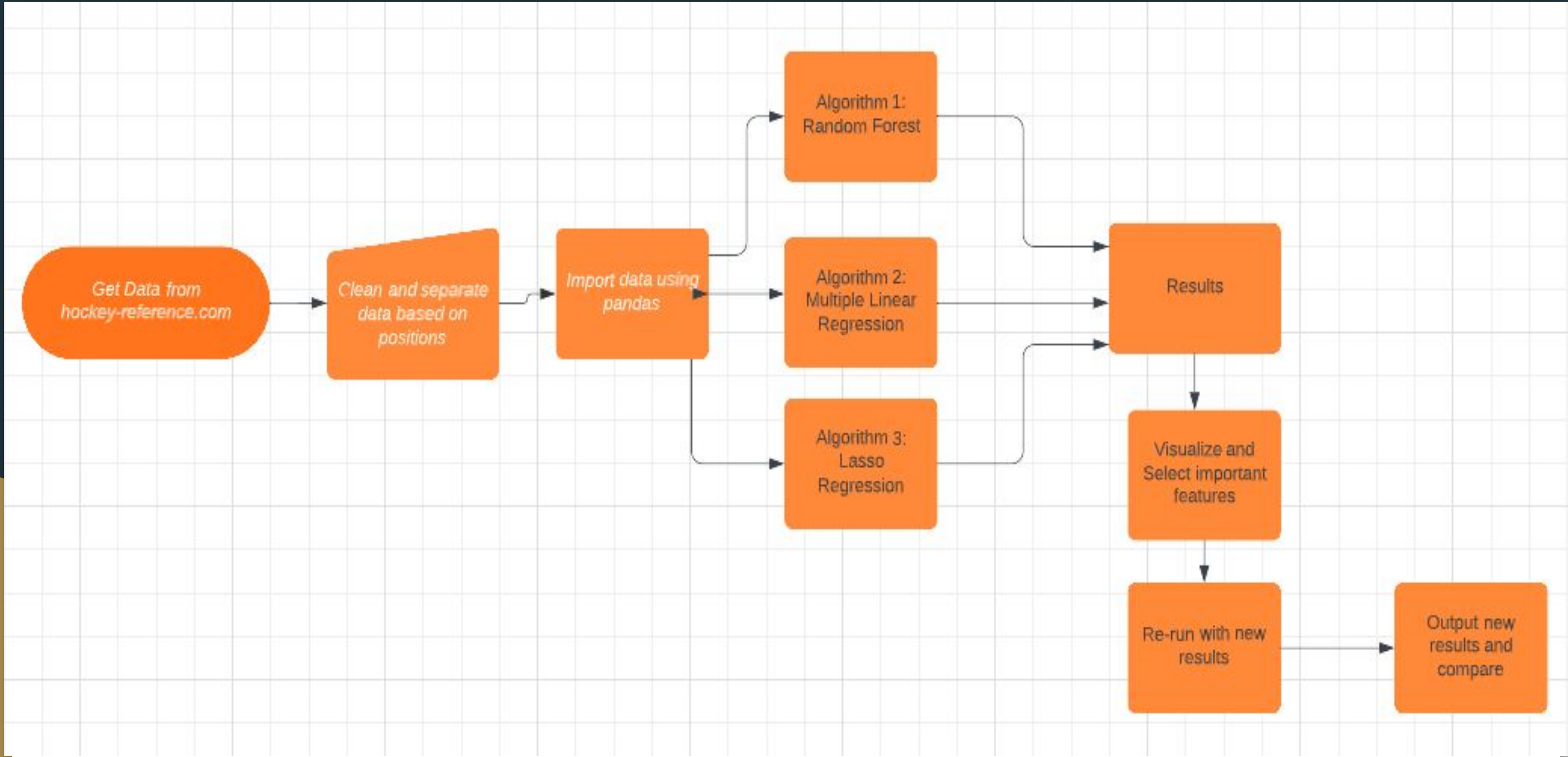
	Age	GP	G	PLMI	CAPHIT
count	222.000000	222.000000	222.000000	222.000000	2.220000e+02
mean	27.279279	68.621622	16.000000	-1.108108	3.117166e+06
std	4.189345	15.586602	11.558625	14.193142	2.770609e+06
min	19.000000	11.000000	0.000000	-38.000000	4.500000e+05
25%	24.000000	60.250000	7.000000	-9.000000	8.941670e+05
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

Data (Wingers 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

The Data (Defenseman 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


Get Data from hockey-reference.com
Clean and separate data based on positions
Import data using pandas
Algorithm 1: Random Forest
Algorithm 2: Multiple Linear Regression
Algorithm 3: Lasso Regression
Results
Visualize and Select important features
Re-run with new results
Output new results and compare

Scoring of the Algorithms

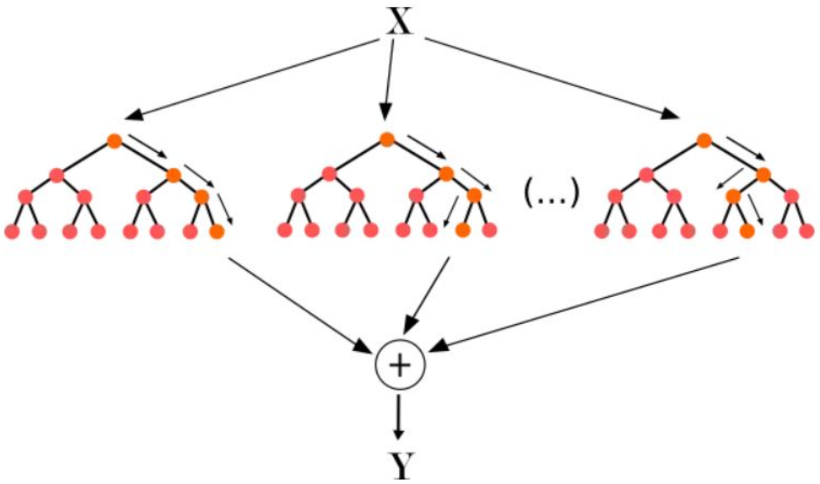
- R-squared:
  - $1 - \frac{(\sum(yi - yp)^2)}{(\sum(yi - ya)^2}$

    Yi = actual cap hit value, Yp = predicted value, Ya = average cap hit
- Mean Absolute Error
  - $\frac{\sum(yi - yp)}{N}$

    Yi = actual cap hit value, Yp = predicted value, N = # of predicted values
- Accuracy Score

Algorithm #1: Random Forest

- Collection of decision trees
- k(final) = (k1 + k2 + k3…+kn)/n
- Binary split at each node
- Highly effective
- Features selected at random
- Final prediction is the average of each decision tree

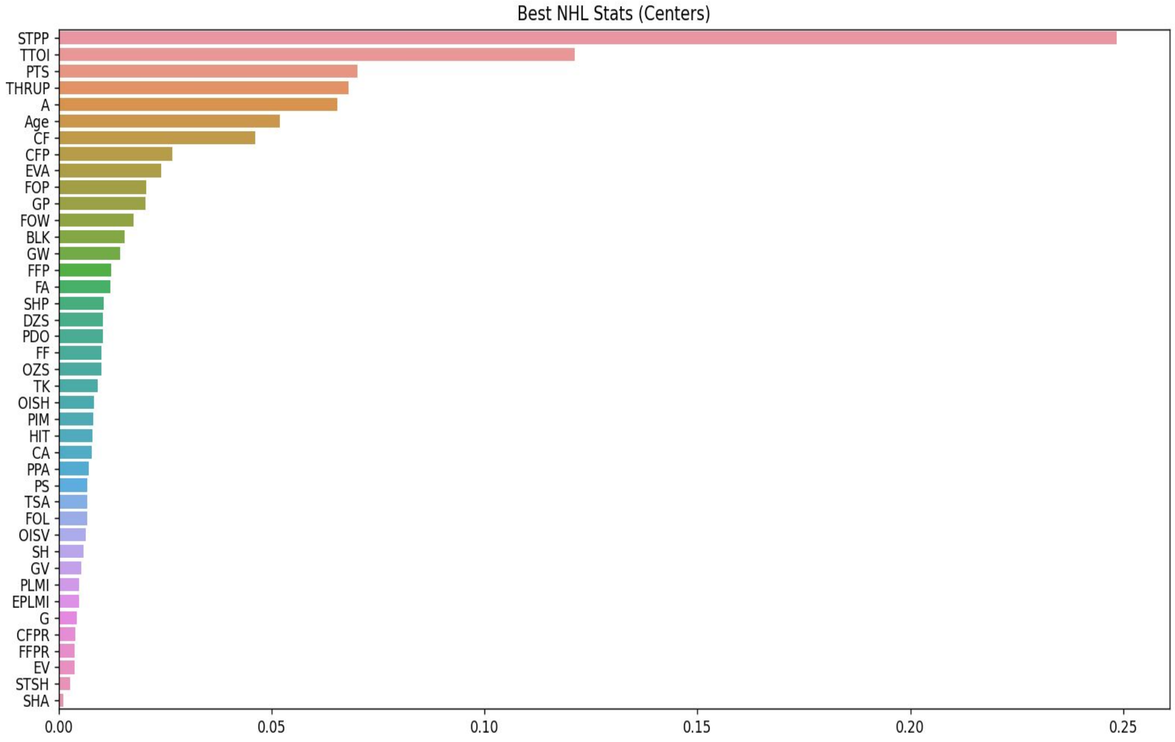


Pros and Cons of Random Forest

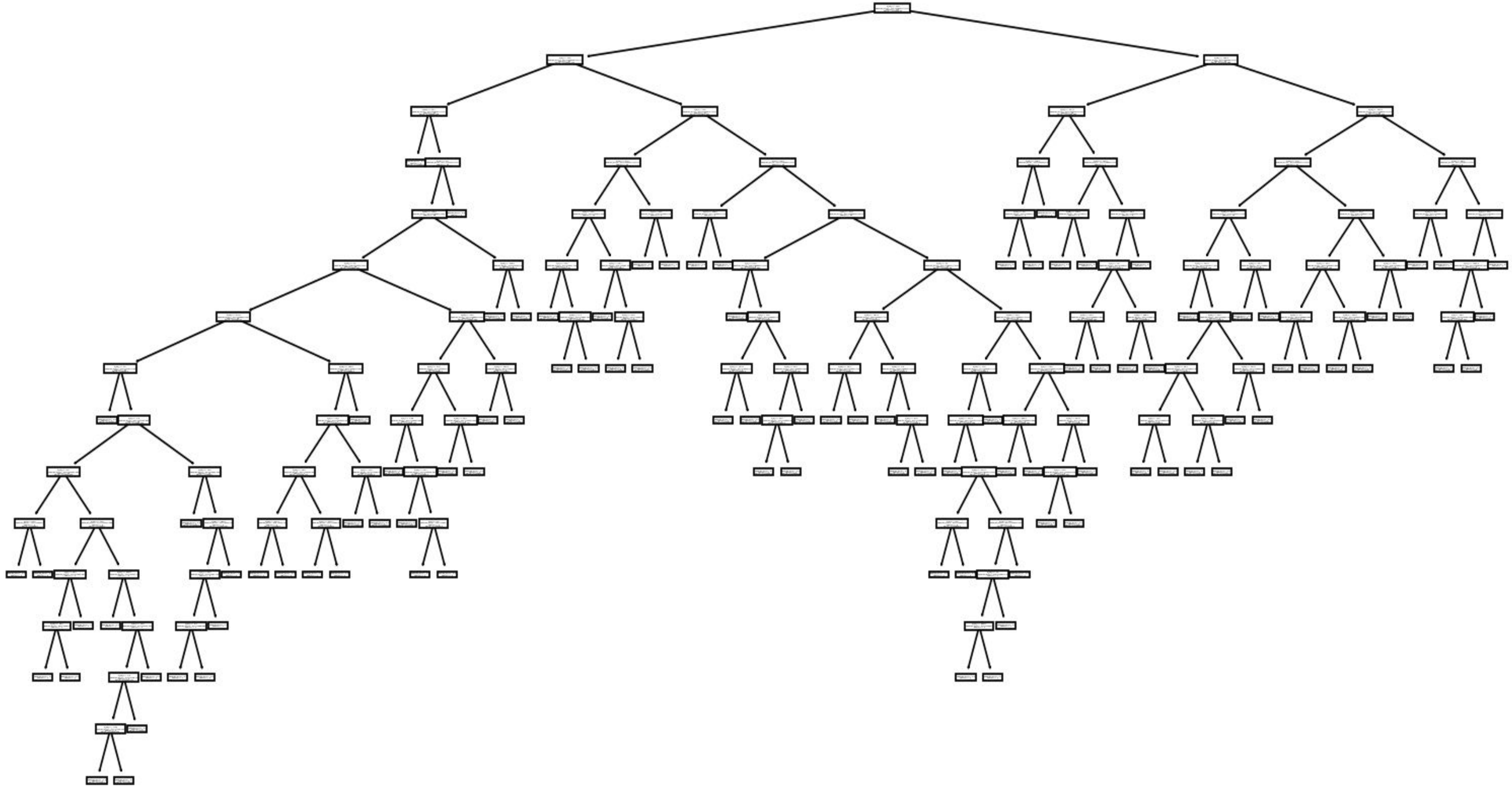
- Good at preventing overfitting
- Provide high accuracy
- Perform feature selection
- Decrease the variance due to averaging out the various trees

- No control over the algorithm
- Can be hard to interpret
- Intense computationally

Algorithm #1: Random Forest (Cont)



- Feature importance graph
- The higher they rank, the more they contribute to the purity of the tree
- Bottom 2 were removed from consideration after the first run through



Random Forest Results

Random Forest Results

Position	R-Squared Value	0	Mean Absolute error	1	Accuracy Score	2
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Centers	0.63	0.62	1421149	1430719	57.27%	57.52%
Defenseme	0.59	0.6	1322208	1288749	34.16%	34.65%
Wingers	0.45	0.45	1446732	1450383	28.36%	29.14%

Algorithm #2: Multiple Linear Regression

- Uses multiple independent variables to output a prediction
- $Y = B0 + B1X1 + \ldots BnXn + e$

- Relies on correlation between the independent variables and the dependent variables

- Uses the whole data set

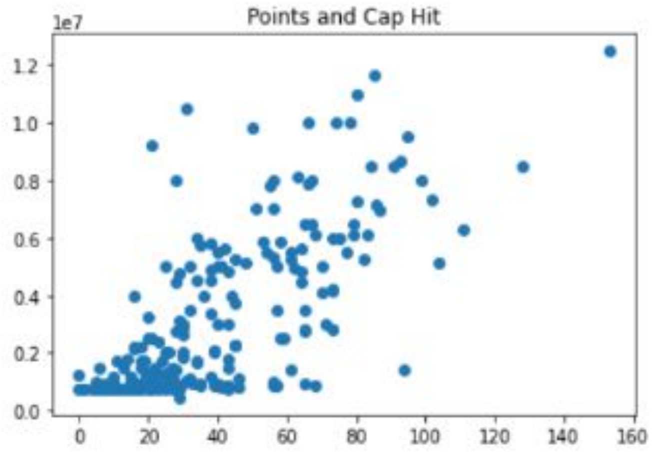
- At the second iteration, stats with a correlation closest to 0 were removed

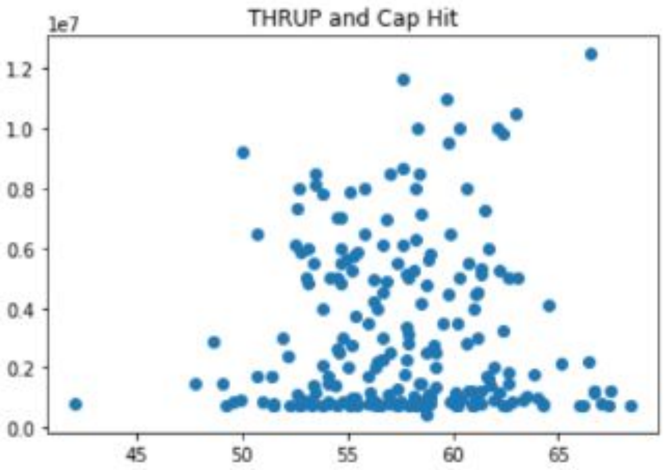
Pros and Cons of Multiple Linear Regression

- Good at determining which factors correlate the most with the dependent variable
- Good at finding relationships between variables
- Good at finding outliers

- Can be ineffective with incomplete data
- Prone to overfitting

Multiple Linear Regression (cont)





Correlation Map

$$\text{Correlation} = \frac{n(\sum(x*y) - (\sum x * \sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y^2))}}$$

	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

Multiple Linear Regression Results

Multiple Linear Regression Results

Position	R-Squared Value	0	Mean Absolute error	1	Accuracy Score	2
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Centers	0.32	0.39	1634126	1617917	13.82%	15.27%
Defenseme	0.58	0.59	1283620	1252264	34.40%	38.36%
Wingers	0.34	0.34	1585959	1523121	10.60%	16.36%

Algorithm #3: Lasso Regression

- Works similar to multiple linear regression
- Each stat has its own correlation coefficient

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Each coefficient is normalized to 0
- Feature selection
- Coefficients that go down to 0 are eliminated

Pros and Cons of Lasso Regression

- Good at preventing overfitting
- Feature selection

- Does not do well when the various features show correlation with each other
- The coefficients are biased

Lasso Regression Results

Lasso Regression Results

Position	R-Squared Value	0	Mean Absolute error	1	Accuracy Score	2
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Centers	0.51	0.63	1513222	1360718	30.85%	37.65%
Defenseme	0.59	0.61	1124918	1136815	19.82%	21.47%
Wingers	-0.4	-0.03	1666787	1440604	-14.14%	5.15%

Overall Conclusions

- Multiple Linear Regression was the least effective of the 3 as the whole data set was used
- Lasso Regression was effective because of feature selection
- Random Forest is the most effective overall due to its flexibility for this project
- Hard to accurately predict salaries
- Multiple factors that can’t be quantified

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/

Lee, Ceshine. “Feature Importance Measures for Tree Models — Part I.” Medium, 8 Sept. 2020, medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3.

Works Cited (cont)

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.

Stephanie. “Lasso Regression: Simple Definition - Statistics How To.” Statistics How To, 27 Apr. 2021, www.statisticshowto.com/lasso-regression.

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

Works Cited (cont)

hockey-reference.com

"NHL Salaries Over the Years." *HockeySkillsTraining*,

www.hockeyskillstraining.com/nhl-salaries-over-the-years/. Accessed 19 Apr. 2023.