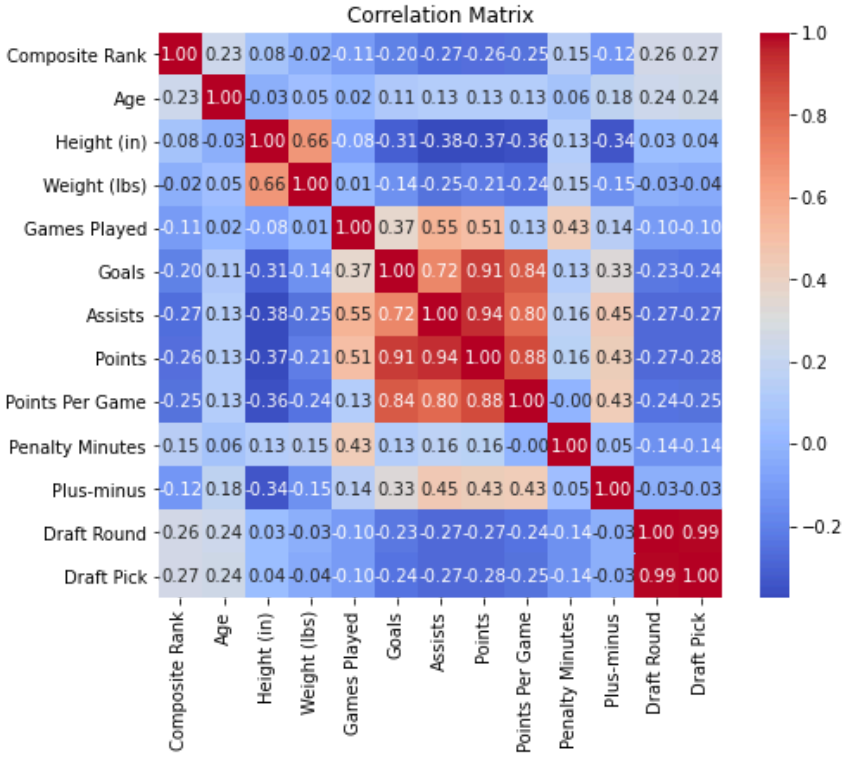
# An Analysis of the 2024 NHL Draft Class Using Classical Statistics and Machine Learning Approaches

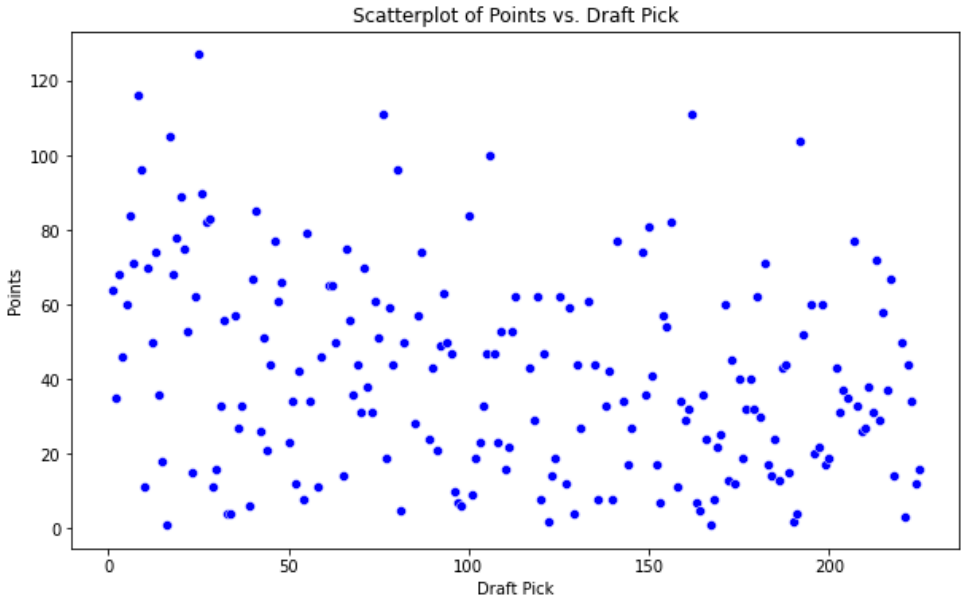
The goal of this project was to see if there were any trends in the most recent NHL draft class. By building models that predict where players will be drafted, one should be able to find the most significant predictors of draft rank or if there are any at all. Note: since goalies are drafted few and far between, and they have different scoring metrics/statistics, I limited my analysis to the other five positions on the ice. To build my models, I used a modified version of The Sound of Hockey's Big Board. This data set combines player statistics from EliteProspects with the rankings of numerous scouts, which are then averaged to create a composite scouting rank. The full code and results for my analysis can be found here.

After separating the data set into position players and goalies, one of the first problems I ran into was that there was missing data in both the penalty minutes and plus-minus categories, especially for players that were lowly ranked. In order to complete my analysis, I opted to fill in the missing data with the median values of the respective categories. I preferred this option to outright removing the penalty minutes and plus-minus variables, as I believed they would be important during the model creation phase.

Before I built any models, I made a correlation matrix, and the results I found were surprising.



Most of the variables have only a low correlation with Draft Pick (where a player was drafted). Goals, Assists, Points, and Points Per Game all had a correlation of ~ -.25, meaning there is a small trend between offensive prowess and Draft Pick. Correlation is negative here because a smaller number for Draft Pick means they were drafted earlier. In contrast, Composite Rank and Age had a correlation of ~ .25. The scatterplot below provides an example as to why these correlations might be lower than expected.




After scaling the variables to all fit a standard normal distribution, I then started with a simple linear regression model on the data, and the results were rather underwhelming.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.158 |
| Model: | OLS | Adj. R-squared: | 0.123 |
| Method: | Least Squares | F-statistic: | 4.493 |
| Date: | Sat, 17 Aug 2024 | Prob (F-statistic): | 5.11e-05 |
| Time: | 13:58:57 | Log-Likelihood: | -1111.5 |
| No. Observations: | 201 | AIC: | 2241. |
| Df Residuals: | 192 | BIC: | 2271. |
| Df Model: | 8 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 112.0249 | 4.401 | 25.452 | 0.000 | 103.343 | 120.706 |
| Composite Rank | 14.9828 | 4.710 | 3.181 | 0.002 | 5.692 | 24.273 |
| Height (in) | 3.3021 | 6.401 | 0.516 | 0.607 | -9.324 | 15.928 |
| Weight (lbs) | -6.4727 | 6.057 | -1.069 | 0.287 | -18.419 | 5.474 |
| Goals | 3.2372 | 7.533 | 0.430 | 0.668 | -11.620 | 18.094 |
| Assists | -8.0670 | 6.614 | -1.220 | 0.224 | -21.113 | 4.979 |
| Points | -3.1303 | 3.777 | -0.829 | 0.408 | -10.579 | 4.319 |
| Points Per Game | -10.5702 | 10.138 | -1.043 | 0.298 | -30.566 | 9.425 |
| Penalty Minutes | -10.0799 | 4.893 | -2.060 | 0.041 | -19.731 | -0.429 |
| Plus-minus | 9.2922 | 5.163 | 1.800 | 0.073 | -0.891 | 19.476 |

| | | | |
|---|---|---|---|
| Omnibus: | 24.216 | Durbin-Watson: | 0.305 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 7.153 |
| Skew: | -0.015 | Prob(JB): | 0.0280 |
| Kurtosis: | 2.076 | Cond. No. | 1.42e+16 |

Of all the predictors, only Composite Rank and Penalty Minutes are significant, however Plus-minus does have a p-value close to the ideal 0.05 mark. The R-squared of 0.158 is also a cause for concern, meaning that the model only encapsulated 15.8% of the data. In combination with what we observed with the correlations, this implies that the problem at hand cannot be solved by simple linear means.

Next, I attempted to find more promising insights with non-linear/improved linear machine learning-based regression models. Using the scikit-learn package, I fit ten different models, scored them using mean squared error and R-squared, and also did feature importance analysis on models that permitted it. The results were an improvement on past endeavors, while also confirming some of the findings from the linear regression model.

| | Mean Squared Error | R-squared |
|---|---|---|
| K-Nearest Neighbors | 1313.502439 | 0.621264 |
| Random Forest | 1514.149344 | 0.563409 |
| Gradient Boosting | 1770.216781 | 0.489575 |
| Neural Network | 2975.018575 | 0.142182 |
| Decision Tree | 3165.414634 | 0.087283 |
| Support Vector Regression | 3378.948519 | 0.025712 |
| Bayesian Ridge | 3438.086147 | 0.008660 |
| Elastic Net | 3836.571806 | -0.106239 |
| Lasso Regression | 3910.887622 | -0.127667 |
| Ridge Regression | 4013.458438 | -0.157243 |

I observed significant improvement to R-squared with the K-Nearest Neighbors, Random Forest, and Gradient Boosting Models, while the models most similar to linear regression performed extremely poorly. Of the decently-performing models that were able to have feature importance analysis done, however, there was a noticeable trend.

Random Forest:

| | Feature | Importance |
|---|---|---|
| 0 | Composite Rank | 0.773296 |
| 7 | Penalty Minutes | 0.048501 |
| 8 | Plus-minus | 0.033126 |
| 1 | Height (in) | 0.031708 |
| 2 | Weight (lbs) | 0.031459 |
| 3 | Goals | 0.028168 |
| 6 | Points Per Game | 0.022915 |
| 4 | Assists | 0.016623 |
| 5 | Points | 0.014203 |

Gradient Boosting:

| | Feature | Importance |
|---|---|---|
| 0 | Composite Rank | 0.795326 |
| 7 | Penalty Minutes | 0.042281 |
| 6 | Points Per Game | 0.034979 |
| 1 | Height (in) | 0.031928 |
| 2 | Weight (lbs) | 0.024797 |
| 8 | Plus-minus | 0.024418 |
| 4 | Assists | 0.020469 |
| 3 | Goals | 0.020293 |
| 5 | Points | 0.005510 |

Composite Rank was the only true important feature in both of these models, and when I tested to see how the models performed without Composite Rank, both the Mean Squared Error and the R-squared plummeted in value.

Overall, there are two important takeaways from the analysis I performed: at least within the 2024 NHL draft, where a player is drafted can be modeled and predicted somewhat

accurately, but it is not a linear problem and can only attempt to be solved through machine learning methods. Additionally, how scouts ranked a prospect ended up being the sole most key factor when it came to predict how players are drafted. One important thing to note is that Composite Rank is likely both a compounding variable while also encompassing traits about a prospect not tracked by general game statistics. Offensive prowess makes up part of a scouts rank, but so do “eye test” traits such as their forechecking ability, skating skills, how they enter the zone, and more. On top of that, rankings tend to matter less as the draft goes on, as teams will start to value a prospect’s “high ceiling” more than any skill floor they might have. In order to improve my analysis in the future, I would look to attempt to numericize the “eye test” traits of prospects, take into account a team’s needs, or even assign risk to certain prospects (some KHL players might not make it overseas, for example).