DS 3000 FINAL REPORT

TO: Dr. Eric Gerber

FROM: Aneel Kahlon, Nicholas Labuda, Winnie Chuma, Max Stracher

DATE : 11/29/23

ABSTRACT

In our project to improve NHL draft decisions, we took a close look at the data we gathered. One interesting finding is that most amateur teams only had one goalie drafted over the years, with two standout teams - Brynas IF jr. (Sweden) and Plymouth Whalers (OHL), where five goalies were drafted from each. On the other hand, Brynas IF jr. had less than 10 skaters drafted, Plymouth Whalers had less than 30, and the USA u-18 development team (USDP/USHL) had more than 70.

In addition to this basic analysis, we wanted to see if we could predict how well players would perform based on the info in the table. We used their draft rankings, age, round selection in a multiple regression model to attempt to predict their ratings. Surprisingly, the model performed pretty poorly. The R^2, a measure of how well our predictions account for reality, was quite low at 0.10, and the mean squared error (MSE) was quite high at 43, telling us our model wasn't fitting well.

Digging deeper, we checked each feature on its own. "Overall" had an R^2 of 0.093 and an MSE of 43.92, "age" had an R^2 of -0.012 and an MSE of 48.100, and "round" had an R^2 of 0.082 with an MSE of 44.413. These results, explained more in our analysis, convey that it is incredibly challenging to predict player success in the NHL draft.

INTRODUCTION

Every year teams in NHL pour countless resources into their scouting departments so that they can get an edge in that year's draft by scouting for the best upcoming prospects who will be eligible for the draft. Which teams are the best at converting their draft picks to long-term skilled players, which positions are certain teams better at drafting, which amateur teams produce the best draft prospects? There are many questions that can be asked about how teams have performed historically in the draft, and this analysis could be useful for teams moving forward as they focus their efforts on drafting for specific positions, they may have historically had difficulty with or drafting a more well-rounded group of prospects. Using data about, the round, age and their placement in terms of overall pick, we will attempt to predict the players rating. The rating will serve as the outcome as it is a way to measure how successful one player has been compared to another. To get player ratings we will attempt to combine the extra data about the players statistics across their lifetime in the NHL into a rating metric, however if we are unable to do so in a reliable way, we will turn to outside resources to provide these ratings.

DATA DESCRIPTION

The main data source we used for this project can be found at hockeyrefrence.com. This data set includes the following features:

- year: year of draft
- overall_pick: overall pick player was drafted
- team: team player drafted to
- player: player drafted
- nationality: nationality of player drafted
- position: player position
- age: player age
- to_year: year draft pick played to
- amateur_team: amateur team drafted from
- games_played: total games played by player (non-goalie)
- goals: total goals
- assists: total assists
- points: total points
- plus_minus: plus minus of player
- penalties_minutes: penalties in minutes
- goalie_games_played: goalie games played
- goalie_wins
- goalie_losses
- goalie_ties_overtime: ties plus overtime/shootout losses
- save_percentage
- goals_against_average
- point_shares

The features that we hope to use to predict a player's success include their age, the round they were picked in, and the number overall pick that they were. Additionally, some visual analysis of amateur teams can be found in the accompanying notebook. Alongside these features, we needed some way to rate a player's performance in the league. Although this is doable from the statistics above, we attempted to research how players are typically rated given their statistics and found only vague information about weighted sums of stats. With this information, we decided that without any concrete weights or a known way to regularize the ratings into a given range, another source of ratings would be needed. We attempted to find ratings from EA sports, but they only post their ratings for the top 50 players. Thus, we landed on a csv file that can be found in the GitHub repository referenced in the following article. It should be noted that this file did not contain ratings for goalies and thus from this point on goalies are no longer included in the analysis.

We used data from the drafts between 1995 and 2019 to conduct the analysis. Overall-pick might seem odd as it only available for a player after you draft them; however, we assumed that most players will be picked near their estimated pick, and so this can be used in the place of overall-pick if the model were to be used proactively.

METHOD

In this analysis of NHL draft picks, machine learning tools play a pivotal role in understanding the factors contributing to the long-term success of players. The primary tool of choice is the multiple regression model, employed to predict player success based on the features overall_pick, age and round. This model makes several key assumptions to function effectively, including the independence of observations, which ideally should stem from independent and identically distributed samples. It doesn't assume linearity, although non-linear features must be explicitly made beforehand. Assumptions related to multicollinearity, homoscedasticity, and the normality of residuals should be considered and, when necessary, addressed during data preparation.

The pitfalls of this method are of course choosing which features to use, and at what power/shape. Multiple regression requires choosing specific features as well as specifying their shape (linear, quadratic, etc) before constructing the model. This makes it difficult to be confident that the model is the best it could possibly be. However, this can be overcome with time, trying the model multiple times with different combinations of features and shapes of each feature, checking that the assumptions are met each time.

To determine which features will be used in the model we first investigated a few aspects of the data, and the impacts different features have on the model's outputs. We generated a pair plot of the numerical features of interest in the data, and individual scatter plots of features that showed some promise for linear relationships, these included: age, round, and overall pick all plotted against the 'Rating". After deciding to proceed, we first compared models trained on each feature individually by comparing their single-fold cross validated R^2 values. We also then investigated multiple permutations of the features once again comparing single-fold cross validated R^2 values and mean squared error values. Additionally, to investigate if better models could be created under more controlled conditions, we compared models under which only data from each of the nine rounds was used and compared models across rounds. We also compared models trained only for a singular core position: Center, Defensemen, Left Wing, and Right Wing.

All of the assumptions for each model were met (discussed below), so we decided against varying the shapes of any of the features in the model. This was done as this is mostly chasing a "better performing" model without much basis for it. Much like P-hacking, this practice of changing method without basis to produce a more desirable result is not sound; thus, we avoided it.

In a broader sense, our model seeks to use the data to make a function that takes in the data as an input and produces a rating as an output. This rating is then compared to the actual rating of the player to produce a score. This score is produced for each player, and the combined result is called the R^2 value. Although it should not be confused with accuracy, it lies on the same scale of 0-1, with closer to 1 being better. This value tells you how much of the variation in player ratings can be accounted for using the model. This is an applicable model because we are trying to predict a useful numerical value based on our data, that being the rating of a player. This rating is directly applicable to whether you should draft a player - our goal for the analysis.

RESULTS

Two of the models that showed promise initially were the models trained on the "overall pick" and the "round". For the model trained on "overall pick" the R^2 was 0.095 and the MSE was 44.69. Below the plot of the model can also be seen as well as the plots of the residuals and the cumulative probability plot to check assumptions.

Figure 1.1 Overall vs. Rating Linear Regression Model

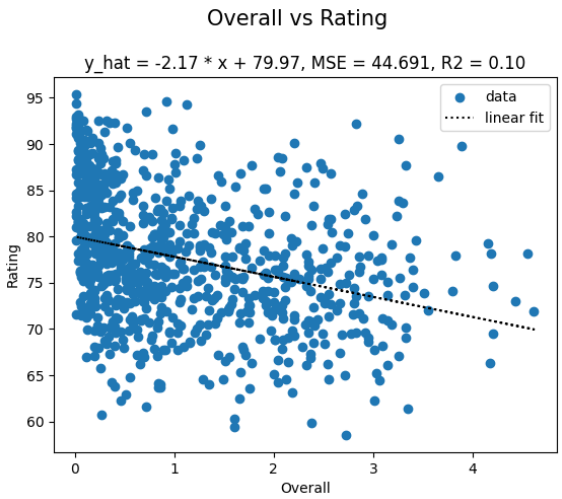
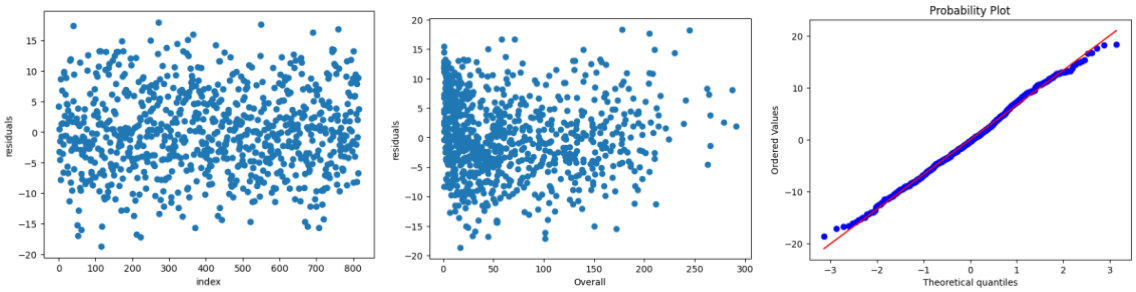


Figure 1.1 Assumption Plots Model 1



For the model trained on "round" the R^2 was 0.086 and the MSE was 45.12. Below the plot of the model can also be seen as well as the plots of the residuals and the cumulative probability plot to check assumptions.

Figure 2.1 Round vs. Rating Linear Regression Model

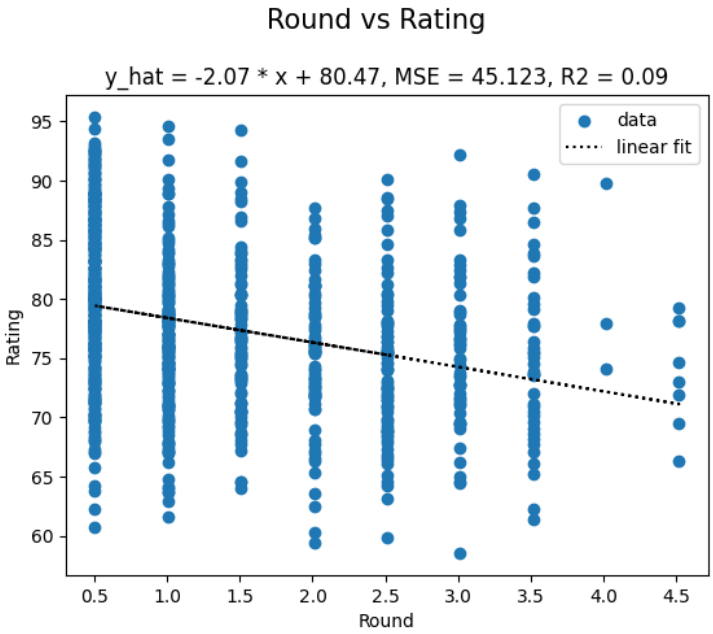
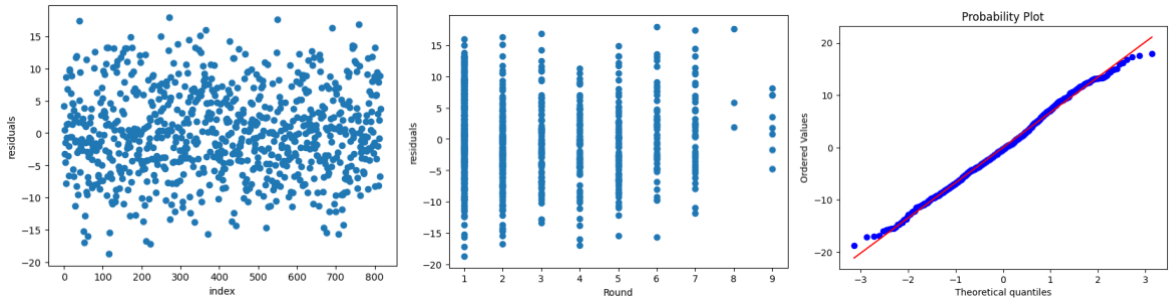
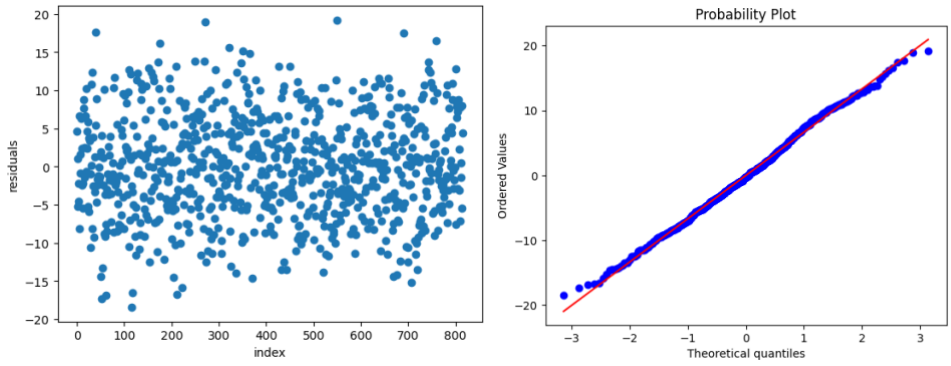


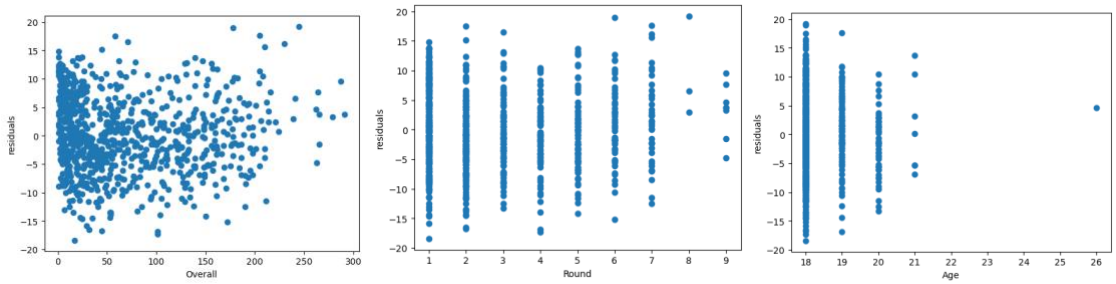
Figure 2.1 Assumption Plots Model 2



For the model trained on "round", "age", and "overall pick" the R^2 was 0.102 and the MSE was 44.35. Below the plots of the residuals and the cumulative probability plot to check assumptions.

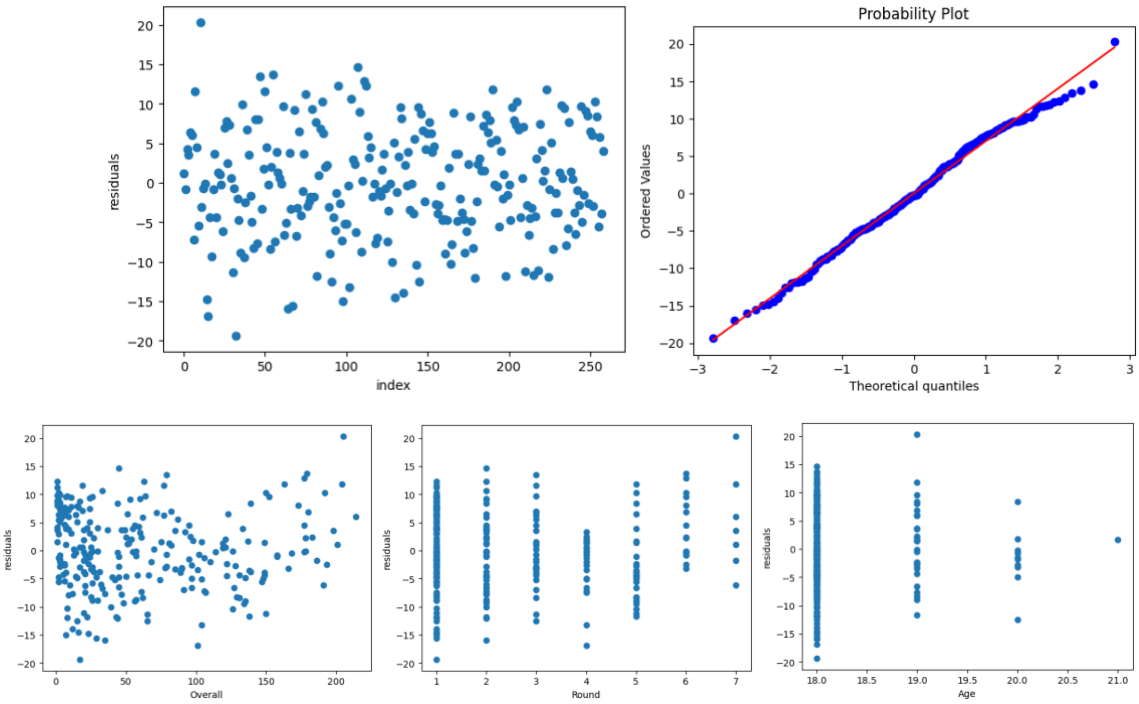
Figure 3.1 Assumptions Plots Model 3





For the model trained which was on only players who were "Centers" and using "round", "age", and "overall pick" the R^2 was 0.221 and the MSE was 48.65. Below the plots of the residuals and the cumulative probability plot to check assumptions.

Figure 4.1 Assumptions Plots Model 4



Discussion

The evaluation of various models provides insights into the factors influencing player success in the NHL draft. The Overall vs Rating graph, based on a single-feature model using the overall pick, suggests a weak relationship (R2 = 0.095, MSE = 44.69). This implies that a player's overall pick alone has limited explanatory power for their draft rating, urging caution for teams relying solely on this metric. To enhance accuracy, exploration of additional features and more complex models is recommended to account for the complexity of factors influencing player ratings beyond the overall pick.

Similarly, the Rounds vs Rating graph indicates limited success in predicting player ratings based solely on the draft round ($R2 = 0.086$, $MSE = 45.12$). The scattered points reveal that a player's success is not consistently determined by the draft round alone, emphasizing the influence of unaccounted factors. Considering additional features or more sophisticated models is advised to improve accuracy, acknowledging the intricate nature of factors shaping player success in the NHL draft.

The MultiFeature Model, incorporating age, round, and overall pick, shows a slight improvement in predictive power ($R2 = 0.102$, $MSE = 44.35$), suggesting that approximately 10.2% of the variability in player ratings can be explained by these features. However, the model's overall explanatory capability remains limited, indicating room for enhancement. Assessing the relative importance of features and considering additional factors is crucial for a nuanced understanding of player success. Future iterations should explore alternative features and adopt more sophisticated modeling techniques to improve accuracy in the NHL draft context.

A more targeted approach is observed in the Center Specific Model, focusing on predictors for the center position. This model exhibits notable improvement ($R2 = 0.221$, $MSE = 48.65$), suggesting that approximately 22.1% of the variability in player ratings for centers can be explained by the included features. The emphasis on position-specific information highlights the importance of tailoring models to unique playing positions. Despite this improvement, the MSE of 48.65 indicates room for further refinement, emphasizing the need to explore additional position-specific features for a comprehensive understanding of player success.

The reliability of the linear regression model is underscored by the assumption plots for residuals vs index, overall vs residuals, and theoretical quantiles. These plots consistently show patterns indicating independence of errors, constant variance, and normal distribution of residuals. While these findings are reassuring, the recommendation for continuous validation and exploration of additional features remains essential for ongoing model refinement. This iterative process ensures that the predictive model remains accurate and valuable for decision-makers in the dynamic context of the NHL draft.