NHL Player Analysis

Sawyer Jacobson

3/13/2020

Introduction

In the hockey world, the National Hockey League is considered the highest professional level of play. As with nearly all professional sports, players are paid ludicrous salaries for the job they perform defined by role and position on the ice. The hockey positions that we look at in this data are forwards (centers, left and right wings) and defensemen. The NHL utilizes a salary cap for teams to prevent a single team with a large budget from signing all of the best players in the league simply because they have the funding to, an issue that is seen more in professional baseball. Therefore it is essential for teams to balance out their budget to properly compensate players based on their performance while staying under the salary cap and balance pay across their roster to attract the best talent while keeping everyone happy. As data analytics in sports becomes increasingly popular and necessary, using statistical tools and tests can allow teams to better determine if players are being utilized properly and, if they are not, giving evidence as to what might be best for them to reach their potential and best benefit the team. Scouts help teams find talent, but their approach barely scrapes the surface on how players should be evaluated as they use a high level of assessment. Statistical analyses will allow for a greater understanding of players from how they play to how they should be compensated and if a player will be a good fit on their roster.

The Data

The data we will be using consists of all NHL player data from the 2018-2019 regular season. The data was collected from the statistics section of https://www.NHL.com by extracting the JSON file that populates the statistics table. Summary statistics for each player over the course of the 2018-19 regular season are included in the dataset. The statistics in our dataset are standard game stats such as goals, assists, points, average time on ice per game, penalty minutes, shifts

per game, shots on goal, hits, etc. as well as player statistics such as birth country, height (in inches), weight (in pounds), birth year, draft status, etc. In total, the data set contains 906 observations with each observation being a player. We obtained NHL player salary data from https://www.hockey-reference.com/friv/current_nhl_salaries.cgi using the `rvest` function to get salaries for 760 players. The player data and salary datasets were inner joined by player name to obtain our final dataset with contained 642 players and 34 total variables. We continued to refine our dataset by filtering out players that played less than 30 total games during the season. This allowed us to analyze more consistent players in the league and eliminate, for example, rookies that only played a few games at the beginning or end of the season. We thought to do this when we noticed an extreme outlier in Ryan Poehling of the Montreal Canadiens. Ryan played in only the last regular season game for the Canadiens after his college season with St. Cloud State ended. The Canadiens did not make playoffs, and Ryan scored 3 goals in as many shots. While impressive, this hat trick skewed the data. The dataset was made more generalizable when outliers such as him were removed. After the data was filtered, we were left with 572 observations that was used to answer our research questions.

Research questions

1. Is there a statistically significant difference in the median player cap hit for forwards and defensemen?

2. Is there a linear (or nonlinear) relationship between time on ice and points per game?

3. Using Random Forest, how accurately can we classify what position a player plays by demographics, including height, weight, cap hit, etc., and game performance?

Brief Exploratory Data Analysis

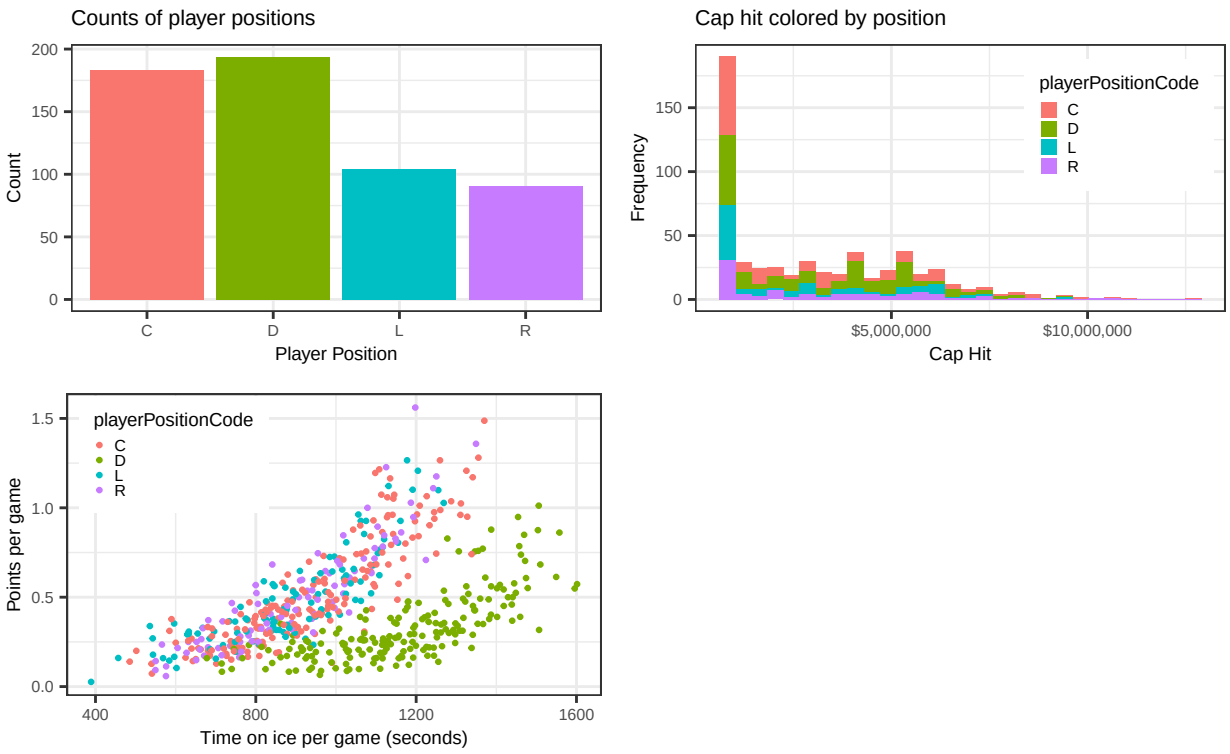

Figure 1: Initial plots of key variables.

The histogram above shows that there is a good distribution in the number of players at each position. The histogram of cap hit shows us that the Cap Hit variable has some pretty severe right skew that we will keep in mind for our analyses. Additionally, in the scatterplot of points per game and time on ice per game, we see a linear relationship. However, we can see two clusters of data with slightly different trends forming, different for forwards and defensemen. Seeing this trend, we will model the second research question separately for these two groups to avoid the issue of heteroscedasticity in linear modeling.

Analysis

Research question #1

From our histogram above, we know the overall spread of cap hit is not normal, but we will check to see if this trend stands when separated into forwards and defensemen. As we can see, our data is has a good deal of right skew, so in this case, it may be better to use the Mann Whitney U test as the normality assumption of the t-Test is clearly broken.

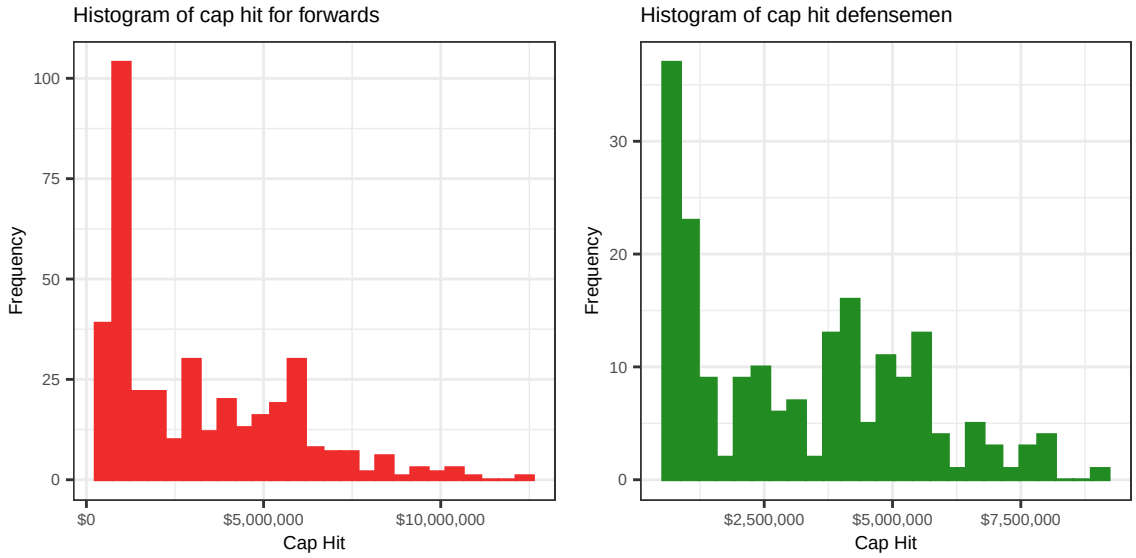

Figure 2: Histograms of player cap hit for both forwards (left), and defensemen (right)

Statistic	P.value	Method
35153.5	0.4189796	Wilcoxon rank sum test with continuity correction

From the results of the Mann Whitney U test, we obtain a test statistic of 3.51535×10^4 and a p-value of 0. Therefore, we fail to reject the null hypothesis and conclude there is not a significant shift in the median cap hit for each position.

Research question #2

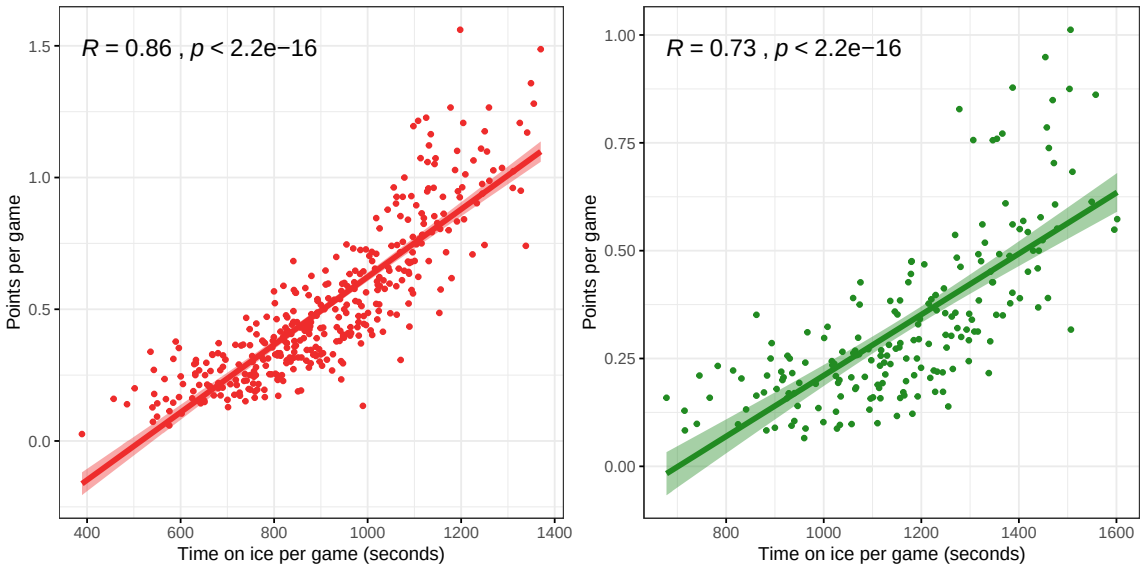

Figure 3: Linear fits for points per game vs time on ice for both forwards (left), and defensemen (right)

From the scatter plots above with regression lines, confidence intervals, and p-values, we see there is a very strong positive correlation between points per game and time on ice per game for both forwards and defensemen. However, we notice there are outliers on the right side of the plots for both forwards and defensemen as well as a slight, but noticeable, nonlinear increase in points per game as time on ice increases for both groups. From these plots and regression lines, we can confirm that there is a strong positive correlation between these 2 variables, but due to the slight nonlinear trend in the points, we cannot conclude this is a strictly linear relationship. A log transformation would potentially increase our model fit. A log transformation of points per game was done in exploration, and a histogram confirms that the normality assumption is better satisfied than with the non-transformed data.

Research question #3

A Random Forest model is a very popular machine learning method that deals with growing a "forest" consisting of a predetermined number of decision trees that have splits for values of each

variable and outputting the mode of the classes for classification, or the mean prediction in a regression case, of the individual trees. Random Forest is very robust as it randomly selects a different subset of variables for each tree as well as bootstrap aggregating to avoid overfitting to the given training data. This method is extremely versatile and can be used in a wide variety of prediction problems. In this paper, we will use Random Forest to answer our last research question of how accurately we can classify players to their respective positions using a Random Forest model. To answer this, we will perform a 70/30 training/test split of our player data. We will train the model on the 70% and test the model performance on the 30%. The variables used in the model are goals, assists, plus inus, cap hit, shifts per game, penalty minutes, weight, height, hits, points per game, and the side the player shoots on.

Model Performance

.metric	.estimator	.estimate
accuracy	multiclass	0.6725146
roc_auc	hand_till	0.7942901

	C	D	L	R
C	35	5	16	9
D	4	62	1	0
L	4	3	7	6
R	5	1	2	11

Table: Accuracy and AUC table (left) and confusion matrix (right, columns are actual).

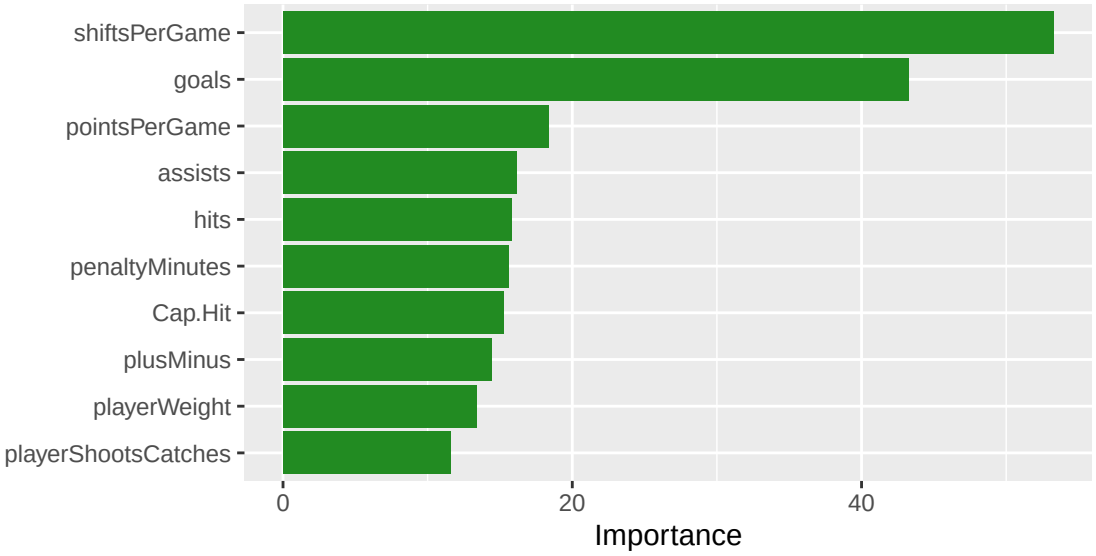

Figure 4: Geni variable importance for the Random Forest model.

From the confusion matrix above, our model falsely predicted the majority of left wingers as centers but faired much better predicting right wingers. However nearly half of them were classified as centers. The model does a very good job predicting centers and defensemen, and the overall model has an accuracy of 0.673 and an AUC of 0.794. Given we have 4 positions players can be classified into and limited training data, the model does an admirable job. The variable importance plot shows which predictors have the most influence in the model predictions. We can see that shifts per game and goals are by far the most important predictors. Intuitively this makes sense because on average, defensemen play more minutes in a game which leads to more shifts, and on average forwards score more goals that defensemen.

Conclusion

Altogether we were able make some interesting findings from our analyses. There is not a statistically significant difference in median cap hit between forwards and defensemen. There is a very strong positive correlation between points per game and time on ice, but this is not a perfectly linear relationship. This relationship appears to be slightly exponential in nature, and, in future analysis, this relationship would be explored using a log transformation. Finally, using a Random Forest model, we were able to predict what position a player is based on certain variables with

decent accuracy. In the game of hockey, there is a lot more to the game than just numbers. There is a significant psychological aspect to the game, straight luck, injuries, etc. The insights gained from this type of analysis would be beneficial when combined with traditional methods of hockey scouting to put to the best product on the ice.