NHL Analytics

Predicting Player Salary

Sawyer Jacobson

4/28/2021

Introduction

Of the most popular sports in the United States, ice hockey is unique in both equipment requirements and the venue where games are played. Unfortunately, some of these requirements can act as a form of gatekeeping for children growing up and not being able to pick the sport up as easily for recreation as baseball or football thus limiting the sports overall growth and popularity. Despite the game being exciting and fast paced, the general fanbase is lacking partially due to these factors. As a result, the National Hockey League (NHL), the sport's professional league akin to football's NFL, sees lower attendance and profits than most of the other major professional sports leagues in the US. Some logistical factors attribute to this as well such as limited seating in venues, but fan interest and excitement is the primary factor. However, in the past 2 decades, the NHL has seen a surge in interest with exciting generational talent, such as Sidney Crosby and Alexander Ovechkin, entering the league and performing to increase the overall quality of product produced by the league night after night. League and team profits increased, and with that, the Salary Cap, the limiting budget set in place by the league as a way to ensure a level playing field amongst all teams, has increased as well thus raising the ceiling for what NHL players can make. A player wants to make as much money as they can while they are in the league, but player performance and their team's current status in regard to the salary cap are taken into effect to determine both what the player is worth and what the team can afford. Player contracts are often multi-year with a different salary value each year, and a player's impact on the team's salary cap is known as their Cap Hit, calculated as the average annual value of their contract. Therefore, a player's Cap Hit will not change over the course of the contract even though the contract may be arranged to have a decreasing salary value over its life as the player gets older and sees a potential decrease in athletic performance. Therefore, using data science to predict a player's salary would be useful both to the player to know what they're worth based on their performance and to the team to assist in team building strategies that are ultimately limited by the salary cap.

Dataset Description/Preparation

Compared to other professional sports, the NHL has not seen as much work or research into analytics and thus does not have as much easily accessed data. A significant portion of this project went towards collecting these data. The data used for this project was collected from 2 separate places. Player statistics data was scraped using the NHL's statsAPI[1]. The API stores the

[1] https://statsapi.web.nhl.com/api/v1/expands

data in JSON format, and, while there is a large amount of data contained there, the overall interface is somewhat clunky with no official documentation. The only documentation that exists is created by fans such as Drew Hynes[2].

Player data from the 2018-19 and 2019-20 seasons were collected. The list of active teams was first collected followed by their respective rosters for the desired season. The data for each player listed on these rosters was then collected. Goaltenders were excluded from this process. Due to the possibility of trades, a player may be listed for more than one team throughout a given season. These data included general season summary statistics such as goals, assists, points, shots, hits, etc. as well as more detailed statistics such as number of goals scored in the second period and number of goals scored when trailing or leading by 2 goals. In total, 1947 player records with 67 variables were obtained. Several of these were purely descriptive in nature and were removed before analysis. While it is possible to see duplicate players, the statistics collected for each player are the aggregated season data so these extra observations are identical except for the team name. As we are not concerned with what team the player was on or ended the season on, the first observation for each player will be selected for use. A total of 1789 non duplicate observations were left. Due to the nature with which these data were collected, missing data for statistics meant the absence of that particular statistic for that player and was filled with 0. Variables such as time on ice (mm:ss format) were cleaned to be usable in analysis.

The player salary data was collected from sportrac.com[3]. The source of where sportrac.com obtains their salary data is unknown and is thus a limiting factor for this analysis. Only data from the 2018-19 to present is freely available. Since the 2020-21 season is still underway, only data from the 2018-19 and 2019-20 seasons were collected. For these 2 seasons, 1342 observations (excluding goaltenders) were obtained with no missing salary or cap hit data. These variables were cleaned to be used in integer format and scaled down to the millions.

As the salary data was obtained from a different source than the NHL API, a matching player key did not exist and thus the statistics and salary data had to be joined by player name and season. Names commonly abbreviated were recorded differently in each source. The players missing observations from the statistics data when joined had their names manually updated appropriately to match the statistics data. Once the data were merged, 1325 observations were left, 849 forwards and 476 defensemen. The modeling was performed on 3 separate groups: overall, forwards, and defensemen. Each group separated into a 70/30 train/test split for modeling.

[2] https://gitlab.com/dword4/nhlapi

[3] https://www.spotrac.com/nhl/

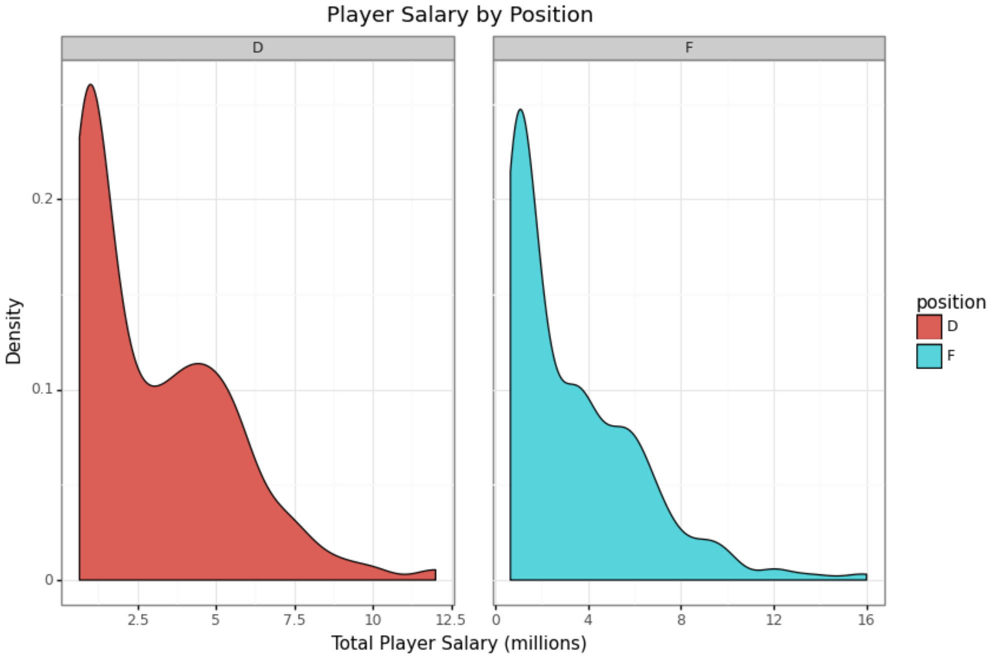


Figure 1: Distribution of player salary separated by position

In Figure 1, we can see that the overall distribution of player salary is relatively consistent between forwards and defensemen. There is a large number of players who make ~$1m that represent rookies on entry level contracts or veterans who sign short contracts on the tail end of their careers. The peak salary for forwards is higher than defensemen as well. Both distributions show significant right skew, but a log transformation (not shown) does not alleviate this and introduces a bimodal distribution.

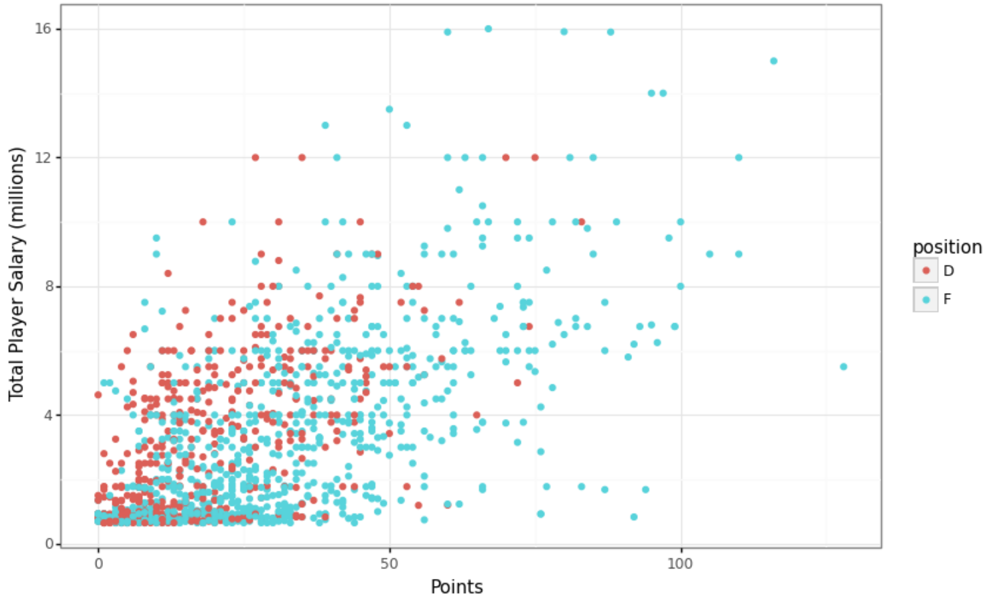


Figure 2: Distribution of player points and total salary displaying a difference in trends based on position.

Figure 2 displays player points plotted against total salary. A differing and expected trend can be observed in that defensemen often have higher salaries despite not being as offensively productive. Based on how the game is structured, this makes sense and gives further evidence that player salaries should be modeled separately for forwards and defensemen as each position is valued differently.

Analysis

As specified above, salary predictions were performed for 3 groups: overall, forwards, and defensemen. The overall salary analysis will not be presented but note that the resulting predictions on the test set performed no better than predicting the mean of the salary variable using mean squared error (MSE) as the error metric. This is an interesting case of Simpson's paradox which occurs when separate underlying trends among groups are hidden by combining all of the groups in an analysis.

The modeling process that follows was performed for each of the forwards and defenseman groups. The predictor variables were hand reduced using domain knowledge of which variables were likely to introduce multicollinearity to the data, confirmed by a heatmap of the variable correlations. One example of this can be seen with the time on ice variables. Overall time on ice is the sum of power play, shorthanded, and even strength time on ice. In these cases, the sum variable was excluded in favor of the more specific variables.

With the data split into train/test sets, Kmeans clustering was used as a form of feature engineering using the traditional game statistics such as games played, goals, assists, plus minus, and shots on goal. Three clusters were used based on the theory that players would be clustered into groups that represent skilled, average, and below average players. These clusters were plotted with the 2 Principal Components, using Principal Component Analysis (PCA), but no noteworthy cluster separations were observed in these 2 dimensions. There was, however, the expected difference between clusters when looking at the median statistic values for each variable. The Kmeans model was used to predict clusters on the test set as well to include in modeling. The numeric variables were standardized, and the remaining categorical columns were encoded using one-hot encoding. Using 5-fold Cross Validation (CV) on the training data, six models were optimized on a group of hyperparameters with the test set MSE recorded. These models included Linear Regression, Random Forest (RF), K-nearest neighbors (KNN), LASSO, Ridge Regression, and Gradient Boosting Model (XGBoost). A final ensemble model was also used to aggregate all of the model predictions with the most overfit model excluded (defined as training error = 0). Furthermore, feature importance was obtained using the LASSO and Random Forest models.

Results

The model predictions were evaluated on their MSE on test set predictions. Figures 3 and 4 below depict the MSE for both the training and the test dataset for forwards respectively. The

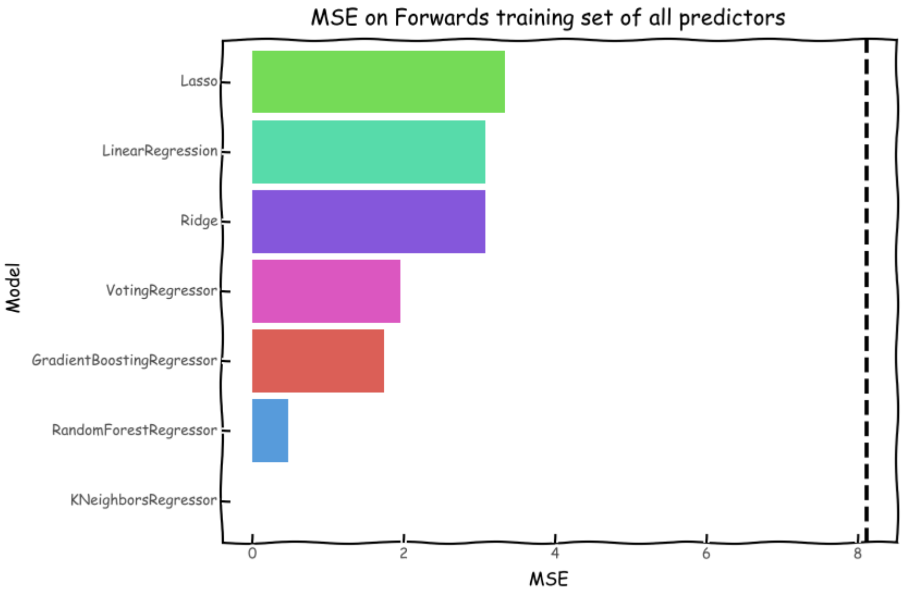


Figure 3: Training set MSE for forwards group for each model.

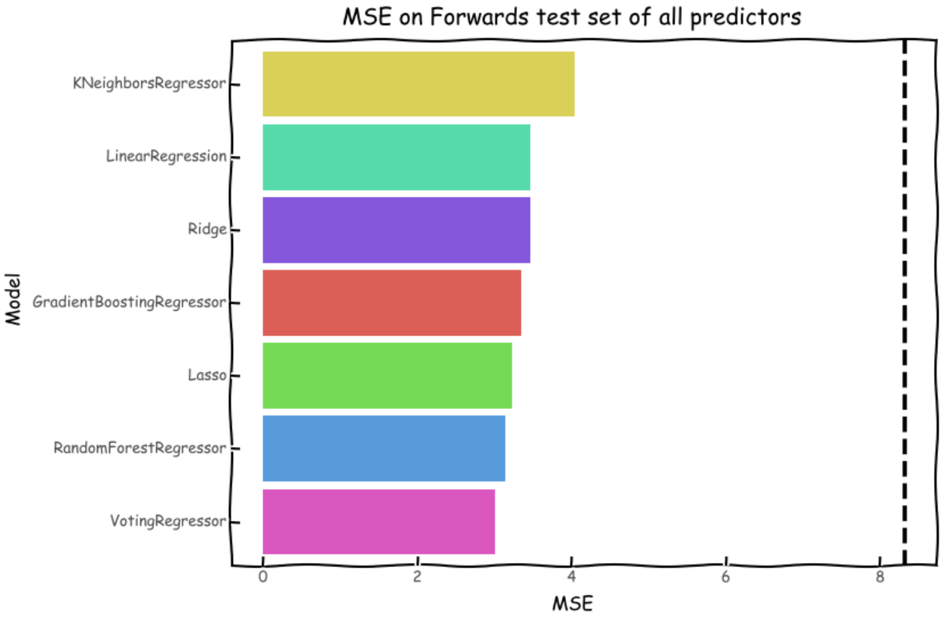


Figure 4: Test set MSE for forwards group for each model.

dashed line represents the MSE on the respective dataset based on predicting the mean for each group. Both of KNN and RF severely overfit the data on the training set. However, despite that, RF was in the top 3 models on the test set for the forwards with the ensemble model performing the best.

Similar modeling results can be seen in Figures 5 and 6 for the defensemen salary predictions with KNN and RF severely overfit on the training data. The LASSO model was able to outperform the ensemble model on the test data for this group by a narrow margin while XGBoost performed almost as poorly as KNN. Removing the RF model from the ensemble would likely cause an increase in performance.

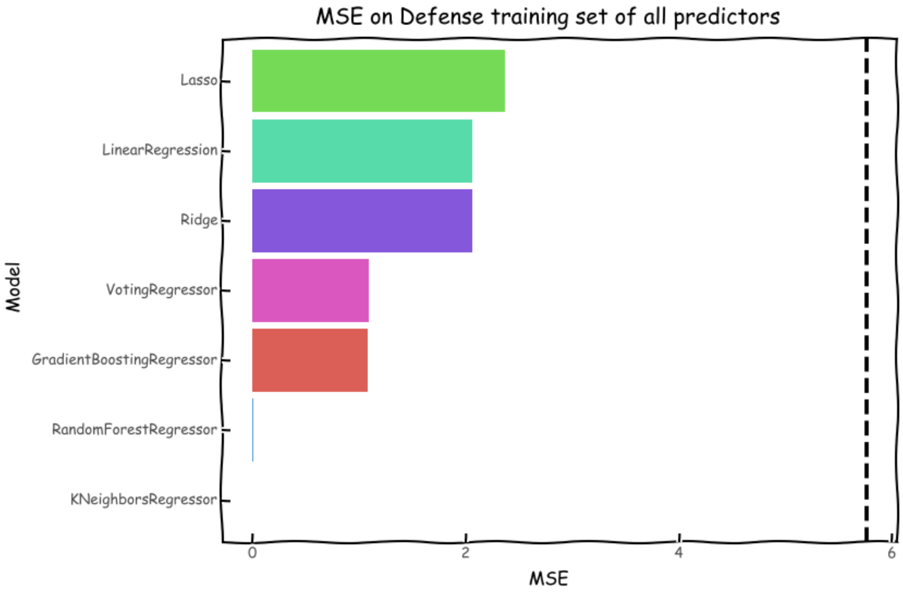


Figure 5: Training set MSE for defensemen group for each model.

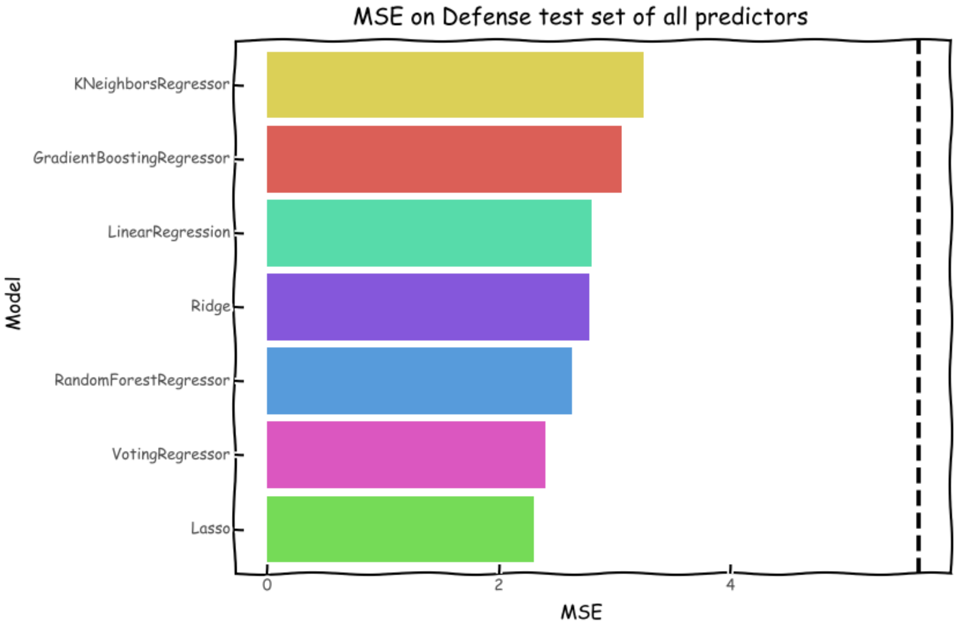
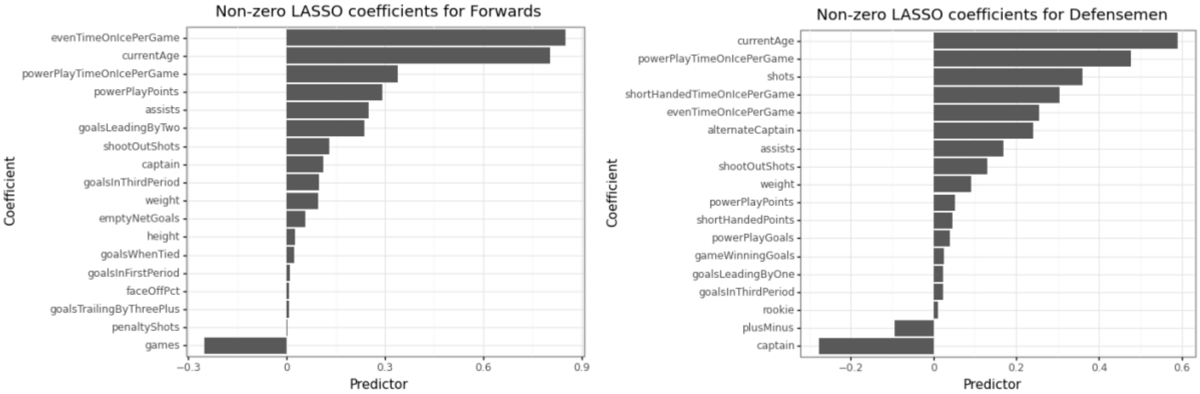


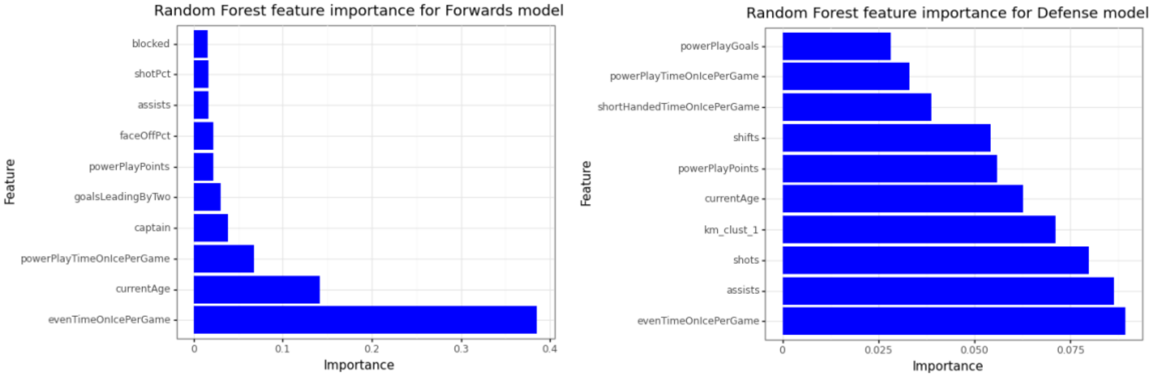
Figure 6: Test set MSE for defensemen group for each model.

Feature importance for the RF and LASSO models can be seen in the figures below. The LASSO coefficients are those that were not reduced to 0 due to the model's inherent feature selection properties while the RF feature importance's were calculated using the Gini index. Figures 7a and 7b below represent the LASSO coefficients for each model. The strongest positive predictors for the forwards model are even strength ice time per game and the players current age while the strongest predictor for the defensemen model is the players current age followed by average powerplay time on ice.



Figures 7a (left) and 7b (right): Non-zero LASSO coefficients for each model.

Feature importance for the RF models can be seen in Figures 8a and 8b. For each model, even strength time on ice was the top feature but carried far more weight for the forward model.



Figures 8a and 8b: Top 10 Random Forest feature for each model.

Discussion

Determining an NHL players salary can be a difficult task. Through the analysis above, we have shown that machine learning can be quite useful but not perfect. The typical MSE for any model was between 2 and 3 million for the defensemen and 3 and 4 million for the forwards. Predicting salaries can be somewhat difficult as they are influenced by more than just performance. For example, players in their first few years in the league are restricted to an entry level contract that has a cap at $875k. After this contract expires, then the player can sign a standard contract or extension that does not have a limitation. Additionally, it is common for older players in the league to sign 1-year, low worth contracts to keep playing even though they may still be performing at a high level. These different contract types are difficult to factor into a model. Moreover, the current financial and salary cap status of a team needs to be taken into account as well. A team might not have the cap space to offer a star player a contract that matches their performance. Another unique case is when a star player will settle for less to allow their team more cap space to retain top talent. An excellent example of this is the Pittsburgh Penguins from the mid 2010s. Sidney Crosby and other star players were due for contract extensions, but each of them took less to keep their team together which resulted in repeat championships in 2016 and 2017.

Overall, machine learning has shown to be a useful tool in predicting NHL player salaries. While there are outside factors that can affect salaries as well, a model such as this could be used by a player and his agent in negotiating a contract renewal/extension. Further work on this project would include acquiring more salary data, which was limited at the time of writing, to potentially reduce model overfitting as well as including more advance hockey analytic statistics. Another approach to salary predictions, given the right data, would be to use a player's statistics from the last season of a contract to predict the starting salary value of their next contract. This would be a more accurate representation of performance translating into salary.