NHL Players Salary Prediction

Big Data Project - Final Report

Alexandre Lacourrege
Masters DSBA
ESSEC-CentraleSupelec
Paris France
b00750454@essec.edu

Deepak Vishal Rajan
Masters DSBA
ESSEC-CentraleSupelec
Paris France
b00746942@essec.edu

Julien Crabié
Masters DSBA
ESSEC-CentraleSupelec
Paris France
b00538002@essec.edu

Ning Wang
Masters DSBA
ESSEC-CentraleSupelec
Paris France
b00740235@essec.edu

Motivation

With sports athletes being extremely well paid, the issue has been put wild attention in the popular press but less in academic research. Professional sport provides us with an opportunity to understand how the salary is explained by the performance of a player. One professional sport attracts our attention. The National Hockey League (NHL), consisting of 31 teams, locating in diverse regions of North America, is a professional Ice Hockey league. With each passing season, fans of NHL keep track of statistics that offer very detailed information of each player as well as individual salaries. Statistical analysis become more and more interesting to find out how hockey is played, and how players are paid. Thus, a salary prediction can enlighten this comprehension. On one hand, a salary prediction model can help players to figure out the essential aspects that are impacting their salaries, and more specifically, what they should do and work on. On the other hand, a salary prediction model helps employers to understand the market value of each player based on his specific performance and contribution to the team.

Objective

The goal of this project is to come up with a salary prediction model based on the available data we have in hand. We want to understand which are the predictors that play an essential role in the salary and whether they could be considered as the most reasonable criterion.

Review of literature

There are research done focusing on individual performance related to player salary and proposing different models. Gomez (2002) conducted a study to examine the relationship between a performance of a player of NHL and his salary. According to his results, individual performance is negatively affected by salary inequality. However, the prediction he made was based on data from 26 NHL teams over five seasons, thus richer data sets are needed.

Fullard (2012) studied two issues. One is to find out whether there is a relation between structure types of salary and predicted NHL team performance. Fullard applied multiple linear regression model and used least square to decrease the error between the regression line and the observed value. The other problem Fullard tackled is the correlation between market value and the player's contribution to the score. However, the results showed no correlation between salary and shootouts, shootout wins or shootout losses.

Then, Peck (2012) studied the salaries of NHL for all positions players of a team. He wanted to find out the important performance variables contributable to the high salary of certain player. Peck stood for the multiple linear regression and said it was sufficient to determine the correlation between the market value of a player and his contribution to the score.

In the study of Louivion (2017) and Pettersson, they used multiple linear regression to understand the relationship between performance and a players' salary. They studied 22 performance-linked variables and the salary as the dependant variable. Their regression model included 12 different covariates and accuracy was 57.4%.

While implementing multiple linear regression, Farrar (1967) pointed out that multicollinearity should also be studied. He used correlation matrix to detect and remove variables that were highly correlated with another from the model. Kendall was the first who proposed Artificial Orthogonalization Method to avoid multicollinearity.

Methodology

In order to tackle our problem, which consist of predicting the salaries of NHL players based on their attributes, we went through the following path:

1. Understanding the variables
2. Missing values treatment
3. Exploratory data analysis
4. Variable transformation
5. Features selection
6. Model selection

1. Understanding the variables

First, it is important to recall that our dataset was split as follows:

- train dataset: 612 players with 154 attributes including Player Salary
- test dataset: 262 players with 153 attributes

Therefore, it seems obvious to start our project by the understanding of the variables. In order to do so, we divided the dataset into 4 parts, parts that were assigned to all 4 team members, for a deeper research about the individual variables pertaining to the technical terms of hockey.

Having a better understanding of the variables enabled us to understand better the correlation matrix, and to make better choices of variable selection between highly correlated variables in the phase of feature selection.

Here are a few examples of the variables that comprised the attributes of the players:

- …
- nzGAPF - Team goals allowed after faceoffs taken in the neutral zone
- nzGFPF - Team goals scored after faceoffs taken in the neutral zone
- NZS - Shifts this player has started with a neutral zone faceoff
- nzSAPF - Team shot attempts allowed after faceoffs taken in the neutral zone
- nzSFPF - Team shot attempts taken after faceoffs taken in the neutral zone
- OCA - Shot attempts allowed (Corsi, SAT) while this player was not on the ice
- OCF - The team's shot attempts (Corsi, SAT) while this player was not on the ice
- ...

2. Missing Value treatment

Our data-sets, both the training one and the testing one, contained some missing values, which needed to be handled in order to enable performing further analysis, such as plots to observe the distribution, plots of correlation matrices to understand the relationship between the variables, etc.

```
sum(is.na(train))

## [1] 426

sum(is.na(test))

## [1] 103
```

When it comes to missing values treatment, several techniques could be used:

- **Deleting the observation**: it is possible to simply delete the values, as sometimes the information might be missing just because it does not apply or exist in a specific case. The code to drop the missing values when encountered is: *na.omit*, which would just drop the entry with the missing values in the data.
  We used this technique for the player "Dan Renouf", who had 53 missing values over a total of 154 variables. Since there were too many nulls in this record, we deleted this observation.

```
train<-train[!(train$First.Name=="Dan" & train$Last.Name=="Renouf"),]
```

- **Deleting the variable**: when encountering some observations / variables with a substantial amount of missing values, it is sometimes advisable to drop the entire attribute or variable, as it does not help in any way for training / assessing the model.
  For instance, for our variable of "Drafted Year", many players had missing values, just because they were simply not chosen during any of the 9 draft rounds. Hence, we decided to drop the column. To compensate for the deletion of this variable, we transformed it to a new variable whose values were between 1 and 9, according to which batch round a player was selected. Therefore, a missing value could be interpreted as "the player was not selected, even after the

9 rounds", that is why we imputed a value 10 to replace the NA. Also, we created the column Experience based on the difference between Draft year and the current year (2016)

```
train$Experience = round(2017 - train$DftYr , 1)
train$Experience[is.na(train$Experience)] <- 0
train$DftRd[is.na(train$DftRd)] <- 10
```

- **Imputing with measures of central tendencies**: In order to prevent the loss of information in the training / assessing of our model, we imputed some values of central measures in order to replace the missing values. Several computations could be done for the replacement, such as *mean*, *mode*, *median.* For many columns in our particular dataset, the most appropriate method was the mean imputation method.
- **Imputing by prediction**: in order to replace missing values, we can use information contained by some related observations that have non-missing values. For instance, K-Nearest-Neighbours could be used

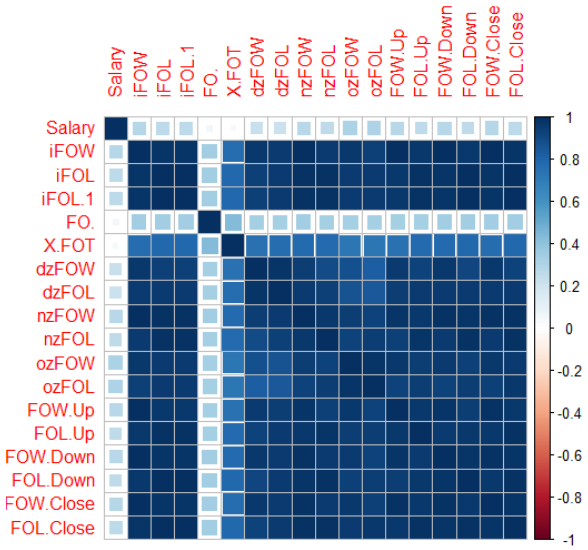
3. Exploratory Data Analysis

Owing to the enormous number of predictor variables in our dataset, the Exploratory Data Analysis was an important part of our project in which we had to invest a lot of time to properly understand our data.

We started with univariate descriptive analysis on each variable. This implies mainly to look at the descriptive parameters of a variable (mean, median, variance, etc.) and to plot histograms for quantitative data in order to visualize its distribution.

We then proceeded with a bivariate analysis which relied mostly on correlation matrices - scatter diagrams were more difficult to interpret considering the number of variables. This preliminary process gave us insights on the relative importance of variables with regards to the other variables.

For example, we had 18 features that were related to face-offs in our starting dataset (a face-off in hockey is a method of beginning play in which two opponents face each other and attempt to gain control of the puck dropped between them by the referee). By constructing correlation matrices and correlation plots, we discovered that 16 of those 18 features were highly correlated together (correlation superior to 0.8). As the correlation of these features with salary were similar we decided to keep only the 2 variables that were the most interpretable (the total face-offs won, and the total face-offs lost by a player).



Apart from this example, we used EDA to find every variable that were highly correlated with other variables. This was done for feature selection.

4. Variables transformation

Some variables of our dataset were not in an appropriate format for any quantitative or computational analysis, in the sense that they were not numerical (but strings instead). For example, the column 'Born' was just the Date of Birth of the players. We reconstructed a new variable called Age which could be directly used inside the model.

```
train$Prefix = ifelse(as.numeric(substr(train$Born, start = 1, stop = 2)) <= 5, 20
, 19)

train$Age = round(age_calc(as.Date(paste(train$Prefix,train$Born, sep = "")),as.Da
te("2016-10-01"), units = 'years'))

train$Prefix = NULL
```

Also, we created a new attribute called Experience which was calculated based on the difference between year 2016 and the date on which the player was drafted first. This variable proved to have a good correlation of around 0.5 with the predictor variable, Salary.

5. Feature Selection

In a massive dataset such as the NHL data where we have more than 150 predictor attributes, it is important that we select only the variables which are necessary and that are important in predicting the target variable. Also, by reducing the number of insignificant variables we can drastically reduce the compute time for training the model. We removed a few variables which were duplicates of other attributes. By manually examining the data, we saw that there were a few columns which were just an aggregated value of multiple other columns, showing a very high correlation too. Hence, we just retained the aggregated value column and removed the other root variables. Also, we eliminated some variables which were highly correlated to other variables in the same dataset.

After the process of manual feature selection, we resorted to a much strict unsupervised feature selection method using random forest model. Once the random forest model had been trained, we obtained the importance of each variable in the model based on the value of Mean Decrease in Gini Impurity; higher value referring to an important variable which shows more variability in the dataset. Thus, we select the required number of features to be included for training the model based on the decreasing order of the variable importance.

6. Model Selection and Training

Once the required features to be included in the model have been selected, we had to choose the optimum model which would give us a good accuracy as well as explain the maximum variability in the given data. Initially we tried building a linear regression model with multiple variables as input. Though we achieved a good value of around 0.7 for the mean R squared value (explains the variance in the model), the error percentage of the model, measured through the Mean Absolute Percentage Error (MAPE) was very high (~ 0.83).

Hence, we chose the random forest regression model using the same variables obtained through the feature elimination and selection techniques. Though the percentage of variance explained through the model was around 63%, it still gave a much better MAPE value of around 0.47. MAPE is a measure of mean error between the predicted and the actual value. Hence, the lower the value of MAPE, the better our model is.

After comparing the two models, we resorted to the random forest regression model to estimate the Salary of the NHL players. We observe this increased accuracy in random forest model because they fit the data better into the model without us having to do much variable transformations. It automatically scales the data and it takes care of outliers too. Though it takes a little bit more computing time to train the model when compared to the linear regression model, it gives us a better fit of the data into the model. This explains us that our data is difficult to be fit into a linear model as it possibly does not show a clear linear trend in its regression curve.

Analysis and Conclusion

The objective of this project was to understand the value of a player by analysing their performance metrics. To do so, we successively worked on data cleaning, visualization and transformation. The part of data preparation process was the most fastidious and time-consuming portion in our project. After meticulous analysis of our dataset, we proceeded with manually removing variables that were highly correlated together. After reviewing different model selection and shrinkage method, we opted for the random forest feature selection model to select the 70 most important variables of our dataset. With those variables the random forest regression model was the best to predict salary. This model's fit shows percent of variance explained and MAPE of 63% and 0.47 respectively.

By analysing the results of our model, we found that the most impactful features regarding a player's value is their experience (number of years playing in NHL), which makes absolute sense. The reason is that professional hockey players, like the vast majority of workers in different professions, very rarely negotiate new contracts with lower salary. So, the longer you have played, the more you will be payed (to some extent of course).

The second most important metric for player value is the proportion of time spent on ice by games played. We note here, that in hockey you have a lot of rotations - the 6 starting players substitute often with players on the bench. This globally means that the more you play per game, the higher your salary will be. It's important to mention here, that this is a metric that "globally" assess the performance of a player rather than focusing on specific statistics such as face-offs won for example. This gives us insights on what to focus on while trying to understand the value of a NHL player.

Our model showed that another very important statistic is the team's shot attempts while the player was on the ice. This highlights the importance a player has on how well his team is performing. This gives valuable information on what explains the value of a hockey player and emphasizes the fact that hockey is a team game.

These three performance metrics gave us a better understanding of what mostly defines the value of a player. Still, given our explained variance score and our MAPE, we cannot say that our model exhaustively explains a player's value.

By looking only at the performance, we omit to take into account other significant information such as the agents of the players, the negotiations phases, the state of the market at the time of the contract signature, etc. We denote more generally from our project that the salary of a player cannot solely be explained by his performance. Yet even so, we can affirmatively express that through our data modelling we have found the performance metrics which have the most impact on the Salary of a NHL player.

References and Existing Works:

[1] Gomez, R. (2002), "Salary compression and team performance: evidence from the National Hockey League. Zeitschrift fu, 72, 203-220.

[2] Fullard, J. (2012). Investigating Player Salaries and Performance in the National Hockey League.

[3] Peck, K. (2012). Salary Determination in the National Hockey League: Restricted, Unrestricted, Forwards, and Defensemen.

[4] Louivion, S., & Pettersson, F. (2017). Analysis of Performance Measures That Affect NBA Salaries.

[5] Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics, 92-107.

[6] McCallum, B. T. (1970). Artificial orthogonalization in regression analysis. The Review of Economics and Statistics, 110-113.