National Hockey League – Players Salary Prediction

Big Data Analytics - Project Proposal

Alexandre Lacourrege, Deepak Vishal Rajan, Julien Crabié, Ning Wang

Motivation and Problem Definition

The NHL is a professional Ice Hockey league in North America, currently comprising 31 teams. The **National Hockey League** is considered to be the most prestigious professional ice hockey league in the world. We all know about the controversies which arise in sports clubs when it comes to deciding the salary of every player in the team. Hence in this use case, we will predict the most appropriate salary for every player in the given dataset. On successful completion of the project, we can standardize the method of deciding player salary. The use case that we will tackle can be vividly seen as a **prediction/scoring** problem.

Dataset

The dataset for this use case is taken from Kaggle, which is found under the name ***predict-nhl-player-salaries***. The dataset features the salaries of 874 NHL players with various other variables for the 2016/2017 season. For every player, there are 151 distinct attributes which help in deciding the Salary to be paid for a particular player. The dataset as such is very well described and we have ample data to successfully build a prediction model. The dataset contains information pertaining to the player details such as name, age, height, weight, country, etc. and also the players' attributes such as faceoffs taken, number of assists, goals scored, average score, shots on target, etc.

Methodology

The use case we are trying to solve can be clearly seen as a prediction problem wherein we will be finding the most deserved salary for any player based on his game skills. Like in any predictive analytics problem, it is important that we carefully choose the variables to be used while building our machine learning model. Prior to solving a Data Science problem, it is important to carry out the following procedures: **Data Preparation, Exploratory Data Analysis, Variable selection** and **Big data model selection**.

Data Preparation

The first and foremost operation of any data science problem involves meticulous preparation of the data so that it will be easier to analyse it at the preceding stages. The NHL dataset in Kaggle has been scraped from the publicly available website http://www.hockeyabstract.com/ . Data cleaning tasks such as *field type conversion*, *missing value treatment*, *outlier treatment* and *normalization* will be performed. While exploring this dataset we found that there are many categorical variables given. Appropriate field conversion and variable transformation might have to be done in order to make the data usable for modelling.

Exploratory Data Analysis

It is important to understand the underlying data in a dataset. Hence, we carry out techniques such as *Univariate Analysis, Bivariate Analysis, Variable transformation* and take summary statistics of the important variables. By doing univariate and bivariate analysis, we can get useful visualizations on the distribution of the data and we can compare each variable with one another to identify patterns among them. We will ascertain the importance of the variable based on its distribution and its direct impact to the player's overall performance in the game. We will also carry out dimensionality reduction techniques to get rid of the variables which are of least significance so that we maximize the variance in our data. This will especially be useful in feature selection.

Variable Selection

In our dataset the number of players is not significantly larger than our number of predictors which may lead to increased variability when we try to fit the least squares model. This would result in overfitting and therefore reflect as bad predictions on future observations. Consequently, we must perform variable selection or feature selection on our dataset to eliminate the above risk. To achieve the best variable selection, we will try different methods and compare which one best fit our data.

- We will first try the **subset selection** method. As the number of predictors is too large to perform the best subset selection approach (p > 40), we'll use the **hybrid approach**. This approach can be considered as a mix between the forward and backward stepwise methods. It most closely replicates the best subset selection approach while being less computational demanding.
- We will then try the **shrinkage** method. The underlying assumption is the same, we need to reduce the variance and we will do so by reducing the flexibility of our model. This would result in an increased bias but could still lead to a better bias-variance-trade-off if the decrease in variance is more important. Here we would use the **Lasso method** for better model interpretation.
- **Principal Component Analysis** can also be performed to achieve dimensionality reduction. This also gives us the most important values in the dataset based on the PC values.

Finally, on comparing these approaches we can choose the most significant variables that best fit our model.

Data model selection

After the required features to be included for modelling is selected, it is important to choose the model which best fits to the data to achieve the required prediction. Since this is a Salary prediction problem based on a set of player attributes, we can affirmatively choose **Multiple Linear Regression** to build our **scoring** model. We might have to try including the variables in different combination to obtain the best fitting model.

Evaluation

In order to know how well our model is performing, we need to assess the accuracy of the estimates. The following steps can be performed during the evaluation phase:

- *K-Folds-Cross-Validation* : we can use K-Folds CV in order to tend to a low variance. This consists of randomly dividing the dataset into K-folds of approximately equal size in order to train the model and then test it. We then compute the average MSE with $CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$
- *Leave-One-Out-Cross-Validation* : as our dataset is not that large ($n = 874$) we can use LOOCV in order to obtain the least possible bias. This method is a special case of K-Folds, where $k = n$. Repeating this approach n times gives n values of MSE $(MSE_1, \ldots, MSE_n)$, with an average equal to $CV = \frac{1}{n} \sum_{i=1}^{n} MSE_i$

References

[1] Ashley Jones (2018). Using NHL player performance metrics to predict salary, Medium. (https://medium.com/@ashcan.jones/can-nhl-player-performance-metrics-help-predict-salary-294dd347fcd3)

[2] Andrew Thomas Fleenor (1999). Predicting National Basketball Association (NBA) Player Salaries. University of Tennessee Honors Thesis Projects. (https://trace.tennessee.edu/cgi/viewcontent.cgi?referer=https://www.google.fr/&httpsredir=1&article=1306&context=utk_chanhonoproj)

[3] Koki Ando (2018). NBA Players' Salary Prediction using linear regression model, rstudio. (https://rstudio-pubs-static.s3.amazonaws.com/371407_e21330910f3c4bd2b6e19440013ea793.html)

[4] Kevin Peck (2012). Salary Determination in the National Hockey League: Restricted, Unrestricted, Forwards, and Defensemen, Western Michigan University Honors Thesis Projects. (https://scholarworks.wmich.edu/cgi/viewcontent.cgi?referer=https://www.google.fr/&httpsredir=1&article=3334&context=honors_theses)

[5] Nate Reed (2016). Using Regression to Predict Baseball Salaries, Github. https://www.linkedin.com/pulse/using-regression-predict-baseball-salaries-nate-reed/

[6] Christopher Gillespie (2018). Predicting Starting Pitcher Salaries, Medium. (https://medium.com/discovering-data-science-a-chronicle/predicting-starting-pitcher-salaries-4b7a4a26cb65)

[7] Yuan He - Predicting Market Value of Soccer Players Using Linear Modeling Techniques, Berkeley research.