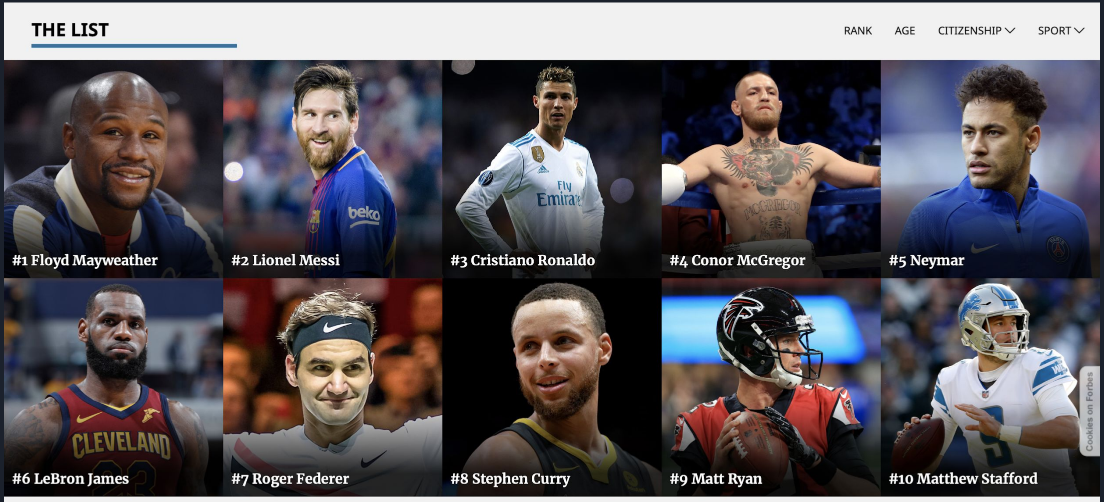
NHL Players Salary Prediction

RAJAN Deepak Vishal, LACOURREGE-MELIN Alexandre, CRABIE Julien, WANG Ning

4th Dec 2018

QUICK INTRODUCTION

At a professional level, certain sports are witnessing extreme variations in player salaries. The question that interested us was the following : **To what extent are these salaries based on historical performance ?**



Our focus : **NHL**

GOALS

- BUILD A MODEL THAT CAN PREDICT NHL PLAYER SALARIES
- UNDERSTAND THE VARIABLES IMPACTING THE SALARY

Methodology

1. Understanding the variables
2. Missing values treatment
3. Exploratory data analysis
4. Variables transformation
5. Feature selection
6. Model selection
7. Evaluation

1. VARIABLE UNDERSTANDING

1. VARIABLE UNDERSTANDING

Example 1

iCF - Shot attempts (Corsi, SAT) taken by this individual

Corsi = shots on goal + missed shots + blocked shots

→ measures how well a player is generating scoring opportunities

→ means a player is keeping the play far away from its own net

Example 2

RelF% - Fenwick percentage relative to his team

Fenwick = (given shots on goal + given missed shots) - (received shots on goal + received missed shots)

→ measures how well a team controls the puck over a game

→ means a player more often in a offensive zone (if positive) than in the negative

2. MISSING VALUES TREATMENT

2. MISSING VALUES TREATMENT

Both our datasets contained missing values, which needed to be treated for the next step of our project.

```
train_sample = head(train, 20)
sum(is.na(train))
```

```
## [1] 426
```

```
sum(is.na(test))
```

```
## [1] 103
```

→ Method chosen for missing values :

- replacement (mean)

```
for(i in 1:ncol(train_num)){
  train_num[is.na(train_num[,i]), i] <- mean(train_num[,i], na.rm = TRUE)

}
```

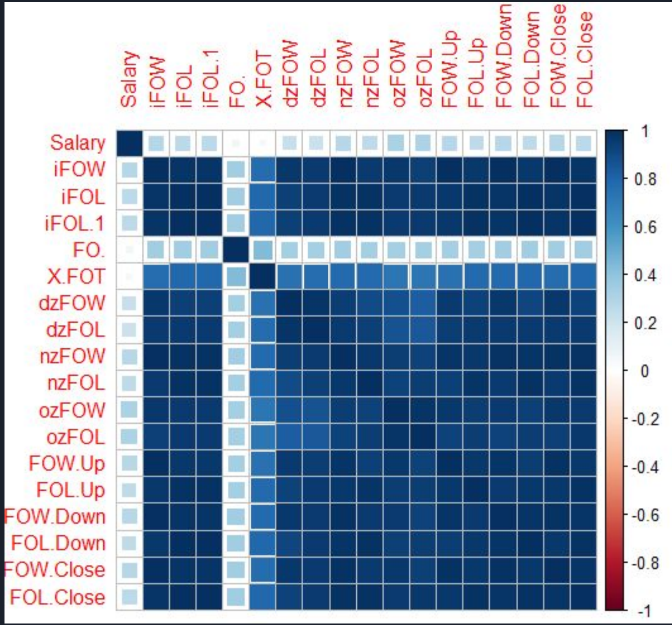
- deleting observation

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

3. EXPLORATORY DATA ANALYSIS

3. EXPLORATORY DATA ANALYSIS

We carried out Univariate and Bivariate Analysis on the raw data to understand more about the variables. After having plotting the correlation matrix, we eliminated the highly correlated variables manually




```
chosen = train[,
c(1,7,8,10:12,15:17,22,24:25,29,32:35,38,40,42,43,44,45,46,49,50,52,55,60,61,
67,68,72,73,85:96,98:107,110,114:116,120,146,147,150:156)]

# chosen = train

tokeep <- which(sapply(chosen,is.numeric))
train_num = chosen[ , tokeep]
```

The same process was repeated for the test dataset

4. VARIABLES TRANSFORMATION

4. VARIABLES TRANSFORMATION (1/2)

Some of our variables were in format that was not proper for our analysis or in a format that could be improved .

```
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.Date("2016-10-01"), units = 'year
s'))

train$Prefix = NULL
```

→ variable conversion : convert “***Born”*** into an appropriate format → numeric “***Age***”

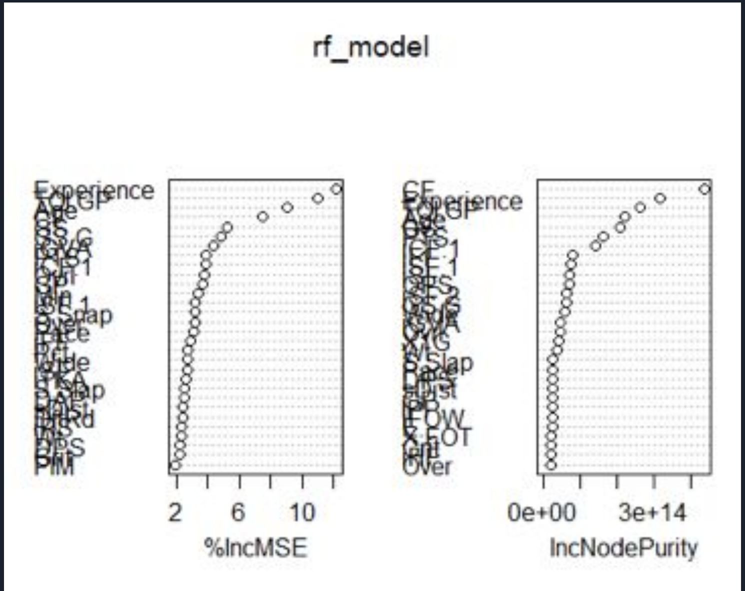
```
test$Experience = round(2017 - test$DftYr , 1)
test$Experience[is.na(test$Experience)] <- 0

test$DftRd[is.na(test$DftRd)] <- 10
test$Ovrl[is.na(test$Ovrl)] <- 0
```

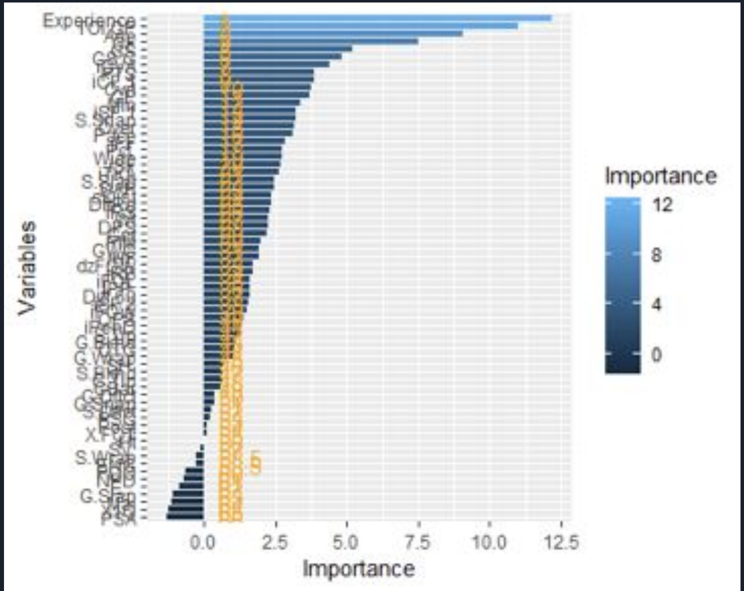
→ variable conversion : convert “***DftYr”*** (draft year) into “***Experience***”.

5. FEATURE SELECTION

5. FEATURE SELECTION

A more strict unsupervised feature selection was applied using random forest model. Higher value refers to a more important variable.





6. MODEL SELECTION

6. MODEL SELECTION

1. **Linear regression model**

Based on the importance of variables, we trained the linear regression model. Although a good value for multiple R^2 obtained, the value of MAPE was high :

```
mean(sm$r.squared)

## [1] 0.6985202
```

```
MAPE(y_pred, y_test$Salary)

## [1] 0.8345729
```

2. **Random forest model**

Comparing the MAPE, random forest model gives better accuracy
→ Our data is difficult to fit into a linear model
→ It does not show clear trend in its regression curve

```
% Var explained: 63.3
```

```
MAPE(y_rf, y_test$Salary)

## [1] 0.5034676
```

7. EVALUATION

7. EVALUATION

Can we explain NHL hockey player's salary by their performance?

Significant metrics

- ❑ Proportion of time spent on ice
- ❑ Experience
- ❑ Team's shot attempts while player on the ice

Limitations

→ Performance doesn't explain everything

→ Significant information not taken into account

APPENDIX

LITERATURE REVIEW

Gomez, 2002	Prediction made based on 26 NHL teams over 5 seasons, richer dataset needed
Fullard, 2012	Multiple linear regression model & least square
Peck, 2012	Confirmed multiple linear regression model & important performance variables found
Louivion, 2017	Multiple regression model
Farrar, 1967	Multicollinearity should also be studied