STAT432_Final_Project

Deadlight

15 十二月, 2019

Abstract

> Football games are always a game that all the world pays great attention to and the football stars sometimes are given much more attention than movie stars. Statistical learning techniques are used to determine if it is possible to effectively predict athletes' wages from data that would be available with many different preditors.

Introduction

Football, or as Americans call it, soccer, is the world's largest spectator sports. Such a large amount of attention would inevitably bring a huge amount of capital into the sport. As a result, athletes are paid very handsomely. For example, Ronaldo's annual income is one of the highest, and he has 93 million US dollars in accounts every year. Messi ranked second with an annual income of 80 million US dollars. His salary and bonus are slightly lower than Ronaldo's, 53 million US dollars, and 27 million US dollars in business income. Famous football stars attract our great attention to their high income.

So it came the questions like "what attributes could earn athlete higher wages", and "how well are athletes across the leagues". To make such predictions, we perform statistical learning techniques on the FIFA19 dataset[1], which includes detailed attributes for every player of last year's latest edition. The game provides a reasonably accurate assessment of player abilities and a tidy-enough dataset to work on.

Methods

A variety of statistical learning techniques were applied. We have used rpart, knn and linear model. We have also used AIC to cut down model size and select the best model. This step is necessary because there exist some obvious meaningless variables in terms of predicting. i.e., the unique ID in the game for a certain player and its jersey number would not be as significant as the player attributes.

Cleaning data

Convert value, wage and weight column to numeric value

The wage, weight and value variables are not in numeric format which is difficult for our use, so we convert them to numeric value.

Dplyr

Find 10 countries with largest maximum and average wage

Nationality	max_wage

Argentina	565
Uruguay	455
Croatia	420
Portugal	405
Spain	380
Belgium	355
Germany	355
Wales	355
Brazil	340
Colombia	315

Nationality	avg_wage
Dominican Republic	71.00000
United Arab Emirates	39.00000
Gabon	26.93333
Egypt	26.15000
Armenia	22.00000
Croatia	21.68254
Central African Rep.	19.00000
Belgium	18.54440
Algeria	18.08333
Brazil	17.81939

Find 10 clubs with largest maximum and average wage

Club	max_wage
FC Barcelona	565
Real Madrid	420
Juventus	405
Manchester City	355
Chelsea	340
FC Bayern M眉nchen	315
Paris Saint-Germain	290
Arsenal	265
Manchester United	260
Liverpool	255

Club	avg_wage
Real Madrid	152.03030
FC Barcelona	146.57576
Juventus	131.68000
Manchester City	113.36364
Manchester United	102.75758
Chelsea	98.45455
Liverpool	87.93939
Tottenham Hotspur	79.48485

FC Bayern M眉nchen	78.82759
Arsenal	78.42424

Modeling

In order to predict the wage, three modeling techniques were considered: linear models, k-nearest neighbors models, and decision tree models.

- Linear models with and without log transformed responses were considered. Various subsets of predictors, with and without interaction terms were explored.
- k-nearest neighbors models were trained using all available predictor variables. The choice of k was chosen using a validation set.
- Decision tree models were trained using all available predictors. The choice of the complexity parameter was chosen using a validation set.

Split to training and testing

Fit linear model using all predictor

The original predictors we choose are player's age, potential, weight, crossing score, finishing the score, Heading Accuracy, Reactions score, balance score and ball control score.

Using AIC to select the best predictor

```
stepAIC(mod1, direction = "both")
```

Linear model

```
## 
## Call:
## lm(formula = Wage_numeric ~ Age + Potential + Weight_numeric + 
##     Crossing + Finishing + HeadingAccuracy + Reactions, data = df_trn)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -39.32  -7.87  -2.42   4.32 498.86 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     -154.53738    2.80741 -55.046  < 2e-16 ***
## Age                0.66062    0.04796  13.774  < 2e-16 ***
## Potential          1.46543    0.03711  39.486  < 2e-16 ***
## Weight_numeric     0.06049    0.01168   5.178 2.27e-07 ***
## Crossing           0.04394    0.01243   3.535 0.000410 ***
## Finishing          0.04051    0.01069   3.789 0.000152 ***
## HeadingAccuracy   -0.02883    0.01055  -2.732 0.006308 ** 
## Reactions          0.49802    0.02814  17.698  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 17.95 on 14327 degrees of freedom
## Multiple R-squared:  0.338,  Adjusted R-squared:  0.3377 
## F-statistic:  1045 on 7 and 14327 DF,  p-value: < 2.2e-16
```

All predictors are significant

KNN model

```
mod_knn_5 = knnreg(Wage_numeric ~ Age + Potential + Weight_numeric + Crossing +
    Finishing + HeadingAccuracy + Reactions,data = df_trn,k = 5)

mod_knn_10 = knnreg(Wage_numeric ~ Age + Potential + Weight_numeric + Crossing
+
    Finishing + HeadingAccuracy + Reactions,data = df_trn , k = 10)

pred_5 = predict(mod_knn_5, df_tst)
knn5_rmse = RMSE(df_tst$Wage_numeric, pred_5)

pred_10 = predict(mod_knn_10, df_tst)
knn10_rmse = RMSE(df_tst$Wage_numeric, pred_10)
```

Decision tree model

```
mod_dt_01 = rpart(Wage_numeric ~ Age + Potential + Weight_numeric + Crossing +
    Finishing + HeadingAccuracy + Reactions,data = df_trn , cp = 0.01)

mod_dt_001 = rpart(Wage_numeric ~ Age + Potential + Weight_numeric + Crossing +
    Finishing + HeadingAccuracy + Reactions,data = df_trn , cp = 0.001)

pred_01 = predict(mod_dt_01, df_tst)
dt_rmse_01 = RMSE(df_tst$Wage_numeric, pred_01)

pred_001 = predict(mod_dt_001, df_tst)
dt_rmse_001 = RMSE(df_tst$Wage_numeric, pred_001)
```

Results

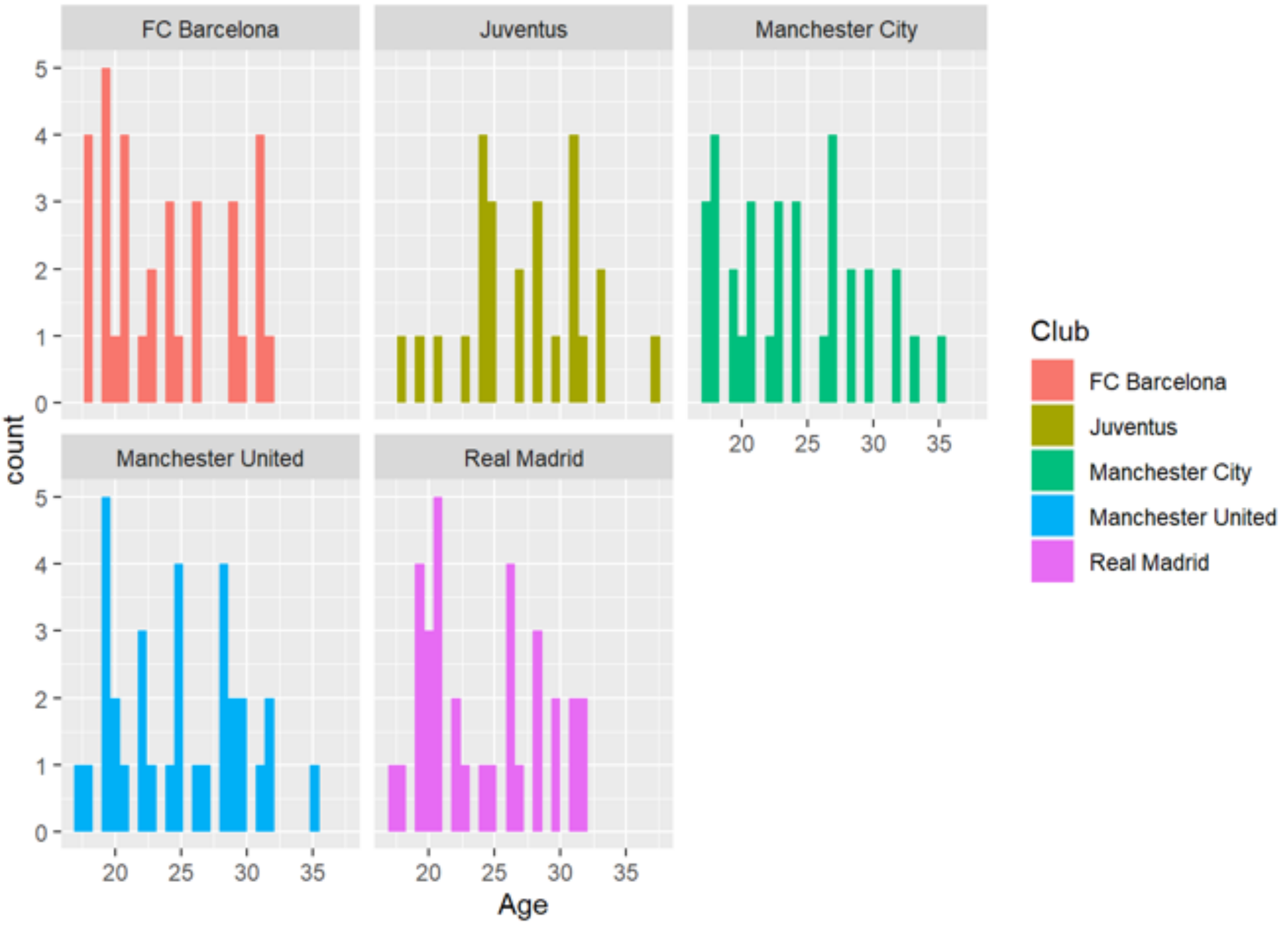
Summary of RMSE for each model

model	RMSE
Linear model	18.42512
KNN model k = 5	13.88752
KNN model k = 10	14.42886

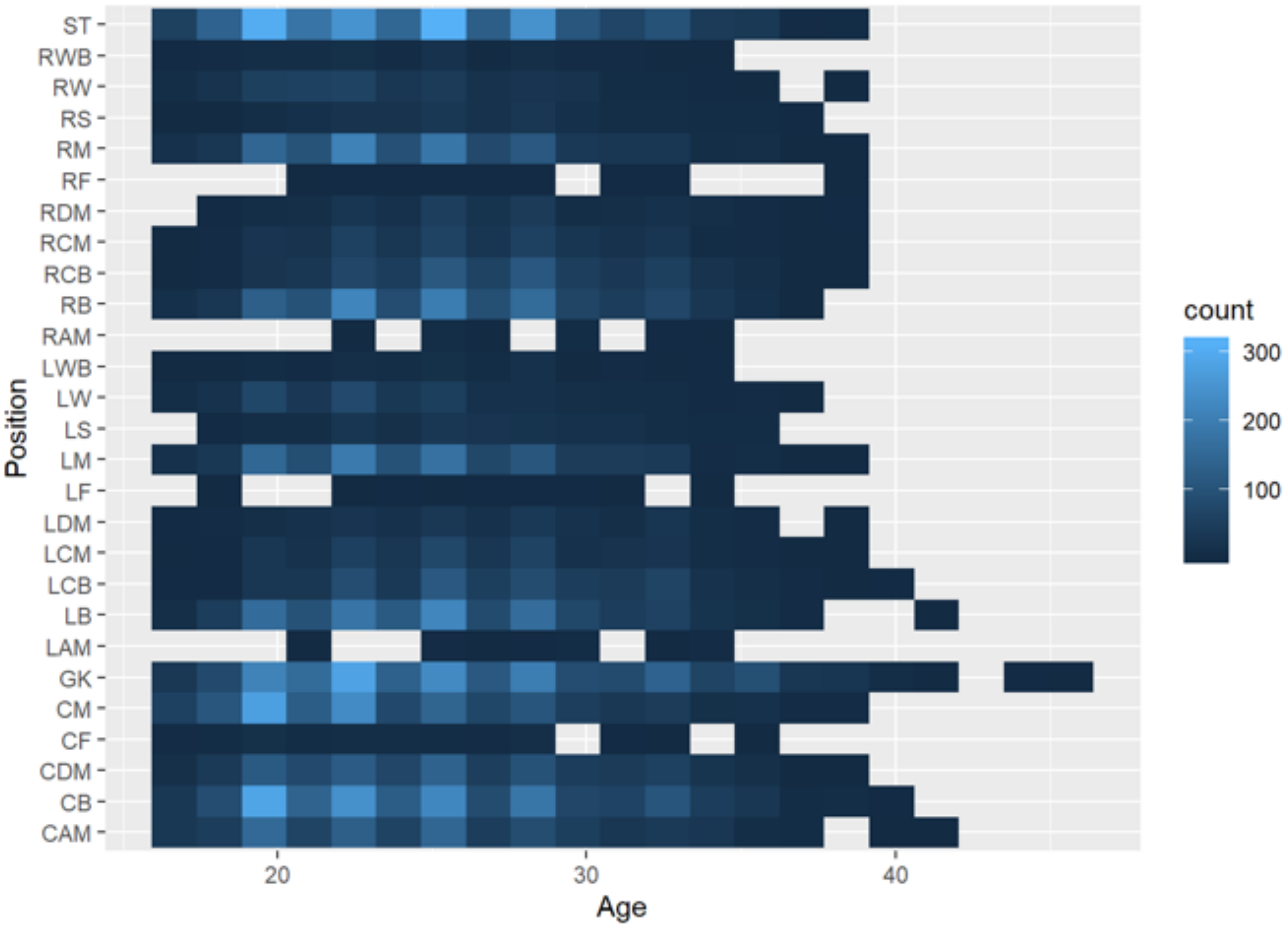
Desicion Tree with cp = 0.01	13.32158
Desicion Tree with cp = 0.001	12.34896

Plot age distribution of five clubs with largest average wage

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Distribusion of position in different age



Discussion

Predicting wage is quite difficult, and there are a lot of factors be considered. Top clubs like Real Madrid, FC Barcelona and Bayern Munich tend to give out the highest wages, and they have a distinguished group of players that are hardly replaceable. Therefore, approximately 100 players earn the highest wages in their respective positions.

On the other hand, even though top athletes are paid astronomical wages, a majority of players are paid less than 50k. Thus it would be difficult for the model to perform ideally across the whole dataset as it essentially covers clubs at all levels. The task is even more difficult as you consider there are leagues in the world, for example, Chinese leagues, that under huge development stages where they give unproportionally high wages to more average players to attract them.

The above notion may be somewhat obscure. A demonstrative example to demonstrate this would be the following pair of players. Consider Willian and Oscar. Both players had played for Chelsea whereas Willian is still playing for Chelsea. They have the same nationality, about the same position on the field and similar international experience for Brazil. However, whereas Willian earns a somewhat modest 120k a week, Oscar earns around a whopping 400k a week. More than three times than that of Willian! This should give the reader some idea on the significance of the difference between the leagues.

On the other hand, I would further elaborate on the ability evaluation of the dataset. FIFA does not provide the most accurate assessment with regards to player abilities. The game functions essentially like NBA2k in the sense that it gathers data using player performance. Nowadays, EA holds the huge scale of collaboration between the game and clubs prior to the release of the new generation. One part of such collaboration is for players to ‘guess’ their FIFA ratings. From this, I suppose some credit can be given on the accuracy of evaluation. In an ideal case, we would obtain some information from soccer statistics site like Squawaka or 442, but this would require techniques like a web crawler and significant data set manipulation techniques that would require more fundamental techniques.

Appendix

Data Dictionary

- ‘Wage_numeric’ - Numeric value of current wqge for each player.
- ‘Value_numeric’ - Numeric value of cuurent value for each player.
- ‘Age’ - Age for each player.
- ‘Potential’ - The potential rating for each player.
- ‘Weight_numeric’ - Numeric value of current weight for each player.
- ‘Crossing’ - Crossing rating on scale of 100.
- ‘Finishing’ - Finishing rating on scale of 100.
- ‘HeadAccuracy’ - Head accuracy rating on scale of 100.
- ‘Reactions’ - Reactions rating on scale of 100.
- ‘Balance’ - Balance rating on scale of 100.
- ‘BallControl’ - Ball control rating on scale of 100.

---

1. Kaggle: FIFA 19 Complete player dataset (https://www.kaggle.com/karangadiya/fifa19)↩