Prediction of Overall

Berkay Taştemel Homework-1

TASK: My task is to create a linear regression model and have it predict the Overall in the FIFA23_official_data dataset.

We use this packages and data set:

```
#install.packages("ggplot2")
#install.packages("car")
#install.packages("tinytex")
#install.packages("readr")
#install.packages("reticulate")
#install.packages("carData")
library(carData)
library(reticulate)
library(readr)
library(tinytex)
library(ggplot2)
library(car)
fıfa23 <- read.csv("FIFA23_official_data.csv", header=TRUE, sep=",", stringsAsFactors=FALS
str(fıfa23)
```

```
'data.frame':   17660 obs. of  29 variables:
 $ ID                      : int  209658 212198 224334 192985 224232 212622 197445 187961 208
 $ Name                    : chr  "L. Goretzka" "Bruno Fernandes" "M. Acuña" "K. De Bruyne"
 $ Age                     : int  27 27 30 31 25 27 30 32 28 28 ...
 $ Photo                   : chr  "https://cdn.sofifa.net/players/209/658/23_60.png" "https:/
 $ Nationality             : chr  "Germany" "Portugal" "Argentina" "Belgium" ...
 $ Flag                    : chr  "https://cdn.sofifa.net/flags/de.png" "https://cdn.sofifa.n
 $ Overall                 : int  87 86 85 91 86 89 86 83 82 88 ...
 $ Potential               : int  88 87 85 91 89 90 86 83 82 88 ...
 $ Club                    : chr  "FC Bayern München" "Manchester United" "Sevilla FC" "Manch
```

```
$ Club.Logo                : chr  "https://cdn.sofifa.net/teams/21/30.png" "https://cdn.sofi
$ Value                    : chr  "€91M" "€78.5M" "€46.5M" "€107.5M" ...
$ Wage                     : chr  "€115K" "€190K" "€46K" "€350K" ...
$ Special                  : int  2312 2305 2303 2303 2296 2283 2277 2273 2271 2262 ...
$ Preferred.Foot           : chr  "Right" "Right" "Left" "Right" ...
$ International.Reputation: num  4 3 2 4 3 4 4 3 3 3 ...
$ Weak.Foot                : num  4 3 3 5 3 4 4 4 4 4 ...
$ Skill.Moves              : num  3 4 3 4 3 3 3 4 3 4 ...
$ Work.Rate                : chr  "High/ Medium" "High/ High" "High/ High" "High/ High" ...
$ Body.Type                : chr  "Unique" "Unique" "Stocky (170-185)" "Unique" ...
$ Real.Face                : chr  "Yes" "Yes" "No" "Yes" ...
$ Position                 : chr  "<span class=\"pos pos28\">SUB" "<span class=\"pos pos15\"
$ Joined                   : chr  "Jul 1, 2018" "Jan 30, 2020" "Sep 14, 2020" "Aug 30, 2015"
$ Loaned.From              : chr  "nan" "nan" "nan" "nan" ...
$ Contract.Valid.Until     : chr  "2026" "2026" "2024" "2025" ...
$ Height                   : chr  "189cm" "179cm" "172cm" "181cm" ...
$ Weight                   : chr  "82kg" "69kg" "69kg" "70kg" ...
$ Release.Clause           : chr  "€157M" "€155M" "€97.7M" "€198.9M" ...
$ Kit.Number               : num  8 8 19 17 23 6 4 15 23 7 ...
$ Best.Overall.Rating      : chr  "nan" "nan" "nan" "nan" ...
```

This data set consists of the data of the players in the Fifa 23 game. Consisting of 17625 observations (football players), this data set includes 29 variables.

SPLITTING THE DATA SET:

```r
set.seed(123)

index <- sample(1 : nrow(fıfa23), round(nrow(fıfa23) * 0.80))

train <- fıfa23[index, ]

test  <- fıfa23[-index, ]
```

Before training this data set, we divide the data set into 2 unequal parts. We also create a seed to get the same observations every time we run these codes.

TRAIN A LINEAR REGRESSION MODEL:

```r
modell <- lm(Overall ~ ID+Age+Potential+Special+International.Reputation+Weak.Foot+Skill.M


summary(modell)
```

```
Call:
lm(formula = Overall ~ ID + Age + Potential + Special + International.Reputation +
    Weak.Foot + Skill.Moves + Kit.Number, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-25.0898  -1.4294   0.2391   1.6781  11.3726

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.865e+01  5.075e-01 -56.453  < 2e-16 ***
ID                        1.854e-05  1.024e-06  18.100  < 2e-16 ***
Age                       1.008e+00  7.037e-03 143.259  < 2e-16 ***
Potential                 8.015e-01  4.140e-03 193.603  < 2e-16 ***
Special                   5.421e-03  1.315e-04  41.222  < 2e-16 ***
International.Reputation  1.404e-01  6.603e-02   2.126 0.033505 *
Weak.Foot                -1.169e-01  3.430e-02  -3.409 0.000655 ***
Skill.Moves              -2.223e-01  4.225e-02  -5.262 1.45e-07 ***
Kit.Number               -1.632e-02  1.154e-03 -14.149  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.489 on 14092 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.9044
F-statistic: 1.668e+04 on 8 and 14092 DF,  p-value: < 2.2e-16
```

We created and trained a model with the lm function, and then we saw the summary of this model with the summary() function.

MEASURING MODEL PERFORMANCE :

```r
prediction.ovrl <- predict(modell, test)

head(prediction.ovrl)
```

```
       8       12       24       28       34       36
84.76128 92.73372 79.60347 85.96783 79.65082 81.37697
```

To test the model, we first find the predicted overalls, for this we used the predict() function.

PERFORMANCE MEASURING WITH RMSE:

```
error <- test$Overall - prediction.ovrl
error <- na.exclude(error)
rmse_model <- sqrt(mean(error ^ 2))
rmse_model
```

```
[1] 2.508654
```

I choose RMSE model because Overall variables is continuous variables. Therefore, the RMSE model is the right choice for my data set.

Overfitting or Underfitting ?

```
rmse_train <- sqrt(mean((modell$residuals) ^ 2))
rmse_test  <- rmse_model
rmse_train - rmse_test
```

```
[1] -0.02073414
```

The result is negative. This result shows us that the performance of the model in the test set is better than the train set.That is, better model performance on the test set may indicate overfitting.

My imaginary player I created in Fifa23 :

```
BerkayTorres <- data.frame( ID <- 151133,
                            Age <- 22,
                            Overall <- 87,
                            Potential <- 93,
                            Special <- 2528,
                            International.Reputation <- 5,
                            Skill.Moves <- 3,
                            Weak.Foot <- 3,
                            Kit.Number <- 9
                             )

BerkayTorres <- predict(modell, BerkayTorres)
BerkayTorres
```

```
       1
84.11277
```

I think the model predicts pretty close.