PREDICTION OF FIFA 23 PLAYERS' VALUES

EMRE AK

3/25/23

In this project we are going to use fifa 23 dataset. With 7 independent value such as (age, overall, potential, weak woot, skill moves, etc.) we try to predict car prices.As you can see below there is libraries which are used for our project and codes.

```r
options(repos = list(CRAN="http://cran.rstudio.com/"))

install.packages("gridExtra")
install.packages("grid")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("tidyverse")
library(tidyverse)
library(ggplot2)
library(dplyr)
library(gridExtra)
library(grid)
library(readr)
```

Our dataset has 17660 row and 29 column.

```r
dim(fifa)
```

```
[1] 17660    29
```

We have just dbl and chr types.

```r
glimpse(fifa)
```

```
Rows: 17,660
Columns: 29
$ ID                          <dbl> 209658, 212198, 224334, 192985, 224232, 212~
$ Name                        <chr> "L. Goretzka", "Bruno Fernandes", "M. Acuña~
$ Age                         <dbl> 27, 27, 30, 31, 25, 27, 30, 32, 28, 28, 26,~
$ Photo                       <chr> "https://cdn.sofifa.net/players/209/658/23_~
$ Nationality                 <chr> "Germany", "Portugal", "Argentina", "Belgiu~
$ Flag                        <chr> "https://cdn.sofifa.net/flags/de.png", "htt~
$ Overall                     <dbl> 87, 86, 85, 91, 86, 89, 86, 83, 82, 88, 84,~
$ Potential                   <dbl> 88, 87, 85, 91, 89, 90, 86, 83, 82, 88, 87,~
$ Club                        <chr> "FC Bayern München", "Manchester United", "~
$ `Club Logo`                 <chr> "https://cdn.sofifa.net/teams/21/30.png", "~
$ Value                       <dbl> 91.0, 78.5, 46.5, 107.5, 89.5, 105.5, 55.5,~
$ Wage                        <dbl> 115, 190, 46, 350, 110, 130, 220, 61, 63, 2~
$ Special                     <dbl> 2312, 2305, 2303, 2303, 2296, 2283, 2277, 2~
$ `Preferred Foot`            <chr> "Right", "Right", "Left", "Right", "Right",~
$ `International Reputation`   <dbl> 4, 3, 2, 4, 3, 4, 4, 3, 3, 3, 2, 4, 3, 1, 3~
$ `Weak Foot`                 <dbl> 4, 3, 3, 5, 3, 4, 4, 4, 4, 4, 4, 4, 5, 3, 4~
$ `Skill Moves`               <dbl> 3, 4, 3, 4, 3, 3, 3, 4, 3, 4, 3, 4, 4, 3, 3~
$ `Work Rate`                 <chr> "High/ Medium", "High/ High", "High/ High",~
$ `Body Type`                 <chr> "Unique", "Unique", "Stocky (170-185)", "Un~
$ `Real Face`                 <chr> "Yes", "Yes", "No", "Yes", "Yes", "Yes", "Y~
$ Position                    <chr> "<span class=\"pos pos28\">SUB", "<span cla~
$ Joined                      <chr> "Jul 1, 2018", "Jan 30, 2020", "Sep 14, 202~
$ `Loaned From`               <chr> "nan", "nan", "nan", "nan", "nan", "nan", "~
$ `Contract Valid Until`      <chr> "2026", "2026", "2024", "2025", "2026", "20~
$ Height                      <chr> "189cm", "179cm", "172cm", "181cm", "172cm"~
$ Weight                      <chr> "82kg", "69kg", "69kg", "70kg", "68kg", "75~
$ `Release Clause`            <chr> "€157M", "€155M", "€97.7M", "€198.9M", "€15~
$ `Kit Number`                <dbl> 8, 8, 19, 17, 23, 6, 4, 15, 23, 7, 7, 10, 2~
$ `Best Overall Rating`       <chr> "nan", "nan", "nan", "nan", "nan", "nan", "~
```

```r
fifa$Age <- as.numeric(gsub("\\.", "", fifa$Age ))
fifa$Overall <- as.numeric(gsub("\\.", "", fifa$Overall))
fifa$Potential <- as.numeric(gsub("\\.", "", fifa$Potential ))
fifa$`Weak Foot`<- as.numeric(gsub("\\.", "", fifa$`Weak Foot`))
fifa$`Skill Moves` <- as.numeric(gsub("\\.", "", fifa$`Skill Moves` ))

set.seed(12)
index <- sample(1:nrow(fifa), round(nrow(fifa)*0.80))
train <- fifa[index, ]
```

```r
test <- fifa[-index, ]
ea <- lm(Value ~ Age + Overall + Potential + `Weak Foot` + `Skill Moves` ,
         data =train, na.action=na.exclude)
summary(ea)
```

```
Call:
lm(formula = Value ~ Age + Overall + Potential + `Weak Foot` +
    `Skill Moves`, data = train, na.action = na.exclude)

Residuals:
   Min     1Q Median     3Q    Max
-447.4 -197.0 -103.7  167.4  795.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   993.8006    33.7487  29.447  < 2e-16 ***
Age             1.3005     0.9864   1.318  0.18738
Overall       -10.2703     0.8206 -12.516  < 2e-16 ***
Potential      -1.2032     0.7898  -1.523  0.12768
`Weak Foot`    -3.0364     3.6145  -0.840  0.40089
`Skill Moves` -11.0307     3.4390  -3.208  0.00134 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 264.7 on 14122 degrees of freedom
Multiple R-squared:  0.1031,    Adjusted R-squared:  0.1028
F-statistic: 324.6 on 5 and 14122 DF,  p-value: < 2.2e-16
```

This Adjusted R-squared is 0.1028, that is not good. That mean is these independent variables explain the target variable %10.28 percently.

```r
y <- predict(ea, test[,-11])
head(y)
```

```
        1         2         3         4         5         6
-12.243507 -18.316263  28.582531  17.101416 -53.032756   6.178921
```

```r
error <- test$Value - y
head(error)
```

```
             1          2          3          4          5          6
     90.743507  95.816263   1.917469  35.898584 168.532756  56.321079
```

```r
rmse_model <- sqrt(mean(error ^ 2))
```

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

```
[1] -4.645968
```

The difference is not positive. That means that the performance of the model is better on test set than train set. that's meaning is maybe overfitting problem.

```r
colnames(fifa)
```

```
 [1] "ID"                       "Name"
 [3] "Age"                      "Photo"
 [5] "Nationality"              "Flag"
 [7] "Overall"                  "Potential"
 [9] "Club"                     "Club Logo"
[11] "Value"                    "Wage"
[13] "Special"                  "Preferred Foot"
[15] "International Reputation"  "Weak Foot"
[17] "Skill Moves"              "Work Rate"
[19] "Body Type"                "Real Face"
[21] "Position"                 "Joined"
[23] "Loaned From"              "Contract Valid Until"
[25] "Height"                   "Weight"
[27] "Release Clause"           "Kit Number"
[29] "Best Overall Rating"
```