Statistical Analysis of Football Matches Data

2024-05-10

Statistical Learning Project aa: 2023/2024

Group members: Flavio Agostini, Niccolò Castellan, Fabio Pimentel

Step 0, defining our goal:

We present a project based on football data.

Step 1, Obtaining data:

As a first step in data acquisition it has been established which parameters could relate to a team's success during a football match. To this end, a questionnaire was submitted to 12 football players with at least 5 years of experience in a semi-pro or pro league. The questionnaire asked them which where the most important factors to win a football match, between:

- Ball possession
- Shots
- Shots on target
- Shots precision (Shots on target / Shots)
- Passage precision
- Km traveled by players
- Quality of the roster (expressed as sum of the values of players)
- Discipline (yellow and red cards)
- Fouls committed and suffered
- Tackles

Participants were free to add suggestions, the most common ones have been: tactics employed, tactical discipline, motivation, referee behavior, field advantage, injuries, level of team-play and experience of the team/roaster. To decide which variables to include in the final dataset, on top of this survey, considerations about data availability and numerical representation of the feature have been made. The features that best suited these criteria and will thus be included in the final dataset are:

Variable	Variable name	Explanation
Field advantage	FieldAdvantage	valued H for "Home", or A for "Away"
Possession	BallPoss	expressed as a percentage
Attempted passages	PassAtt	
Successful passages	PassSucc	
Passage precision	PassPrec	expressed as a percentage
Shots	Shots	
Shots on target	ShotsOnT	
Shots precision	ShotsPrec	expressed as a percentage
Roster quality	RosterQuality	expressed as sum of market value for each player competing in the match, including substitutions

Variable	Variable name	Explanation
Knowledge of the league	Knowledge	expressed as average of matches played in the team's championship by every player competing in the match
Yellow cards	YellowC	
Red cards	RedC	
Fouls committed	FoulsC	
Fouls taken	FoulsT	
Attempted tackles	TacklesAtt	
Successful tackles	TacklesW	
Tackles efficiency	TacklesWRatio	expressed as a percentage
Air duels won	AirDuelW	expressed as a percentage
Attempted dribbles	DribAtt	
Successful dribbles	DribW	
Dribbles efficiency	DribWRation	expressed as a percentage

On top of these features, other four variables, that will serve as target variables, have been gathered: Goals scored, goals taken, outcome of the match and points gained.

10 teams have been chosen for data collection. Of these, 5 teams are the winners of the most prestigious national leagues in Europe. The other 5 have been randomly picked from the same leagues, between teams that didn't perform well enough to qualify for the continental cups but avoided placements that lead to relegation.

"Winners" group:

1. Barclays Permier League (Inghilterra) – Chelsea
2. La Liga (Spagna) – Real Madrid
3. Bundesliga (Germania) – Bayern Munich
4. Ligue 1 (Francia) – AS Monaco
5. Serie A (Italia) – Juventus

"Control" group:

1. Barclays Permier League (Inghilterra) – Everton, pos. 7
2. La Liga (Spagna) – Celta Vigo, pos. 13
3. Bundesliga (Germania) – Eintracht Frankfur, pos. 11
4. Ligue 1 (Francia) – Montpellier, pos. 15
5. Serie A (Italia) – Bologna, pos. 15

Data has been acquired during the year 2018, as part of a BSc thesis project in Physical Education. The data has been collected manually from the Internet. To validate it, for each value a cross-check has been performed between the website whoscored, and the official ESPN website. The only exception to this concerns the parameter "Quality of the roster". To calculate it market values given by transfermarkt have been used. The data has then been saved in a Microsoft Excel spreadsheet for ease of organization and later converted into csv format to be compatible with the R software work environment, where the statistical analysis will be performed.

We import our data and assign it to a dataframe 'df':

```
df <- read.csv('DatasetR_eng_fin.csv', header=TRUE, sep=';',
               fileEncoding="UTF-8")
```

Step 2, Clean and filter data:

Our data is been manually collected and curated so it should already be complete, we nevertheless check for any missing value:

```
anyNA(df)
```

```
## [1] FALSE
```

Using information from "ID" coulumn, we create a new column that identifies as "1" members of the "Winners" group, and as "0" members of the "Control" group:

```
df$Group <- ifelse(grepl("Ca", df$ID), "1",
                   ifelse(grepl("Co", df$ID), "0", NA))
df$Group <- as.integer(df$Group)
```

Let's see what data types we have for our features:

```
str(df)
```

```
## 'data.frame':    372 obs. of  29 variables:
##  $ ID            : chr  "PLCa1" "PLCa2" "PLCa3" "PLCa4" ...
##  $ Match         : chr  "15/08/2016 Chelsea-West Ham United" "20/08/2016 Watford-Chelsea" "27/08/201
##  $ FieldAdvantage: chr  "H" "A" "H" "A" ...
##  $ OutcomeWL     : chr  "W" "W" "W" "D" ...
##  $ PointsWon     : int  3 3 3 1 0 0 3 3 3 3 ...
##  $ Result        : chr  "2 a 1" "1 a 2" "3 a 0" "2 a 2" ...
##  $ GoalsMade     : int  2 2 3 2 1 0 2 3 4 2 ...
##  $ GoalsTaken    : int  1 1 0 2 2 3 0 0 0 0 ...
##  $ BallPoss      : chr  "61,70%" "62,40%" "60,10%" "55,10%" ...
##  $ Shots         : int  16 13 22 28 12 9 22 16 14 13 ...
##  $ ShotsOnT      : int  7 4 10 7 4 2 9 6 6 7 ...
##  $ ShotsPrec     : chr  "43,75%" "30,77%" "45,46%" "25,00%" ...
##  $ PassAtt       : int  480 488 545 416 474 445 460 458 364 325 ...
##  $ PassSucc      : int  556 560 626 488 540 535 554 553 447 409 ...
##  $ PassPrec      : chr  "86,33%" "87,14%" "87,06%" "85,25%" ...
##  $ RosterQuality : int  365500000 378500000 365500000 387500000 422500000 383000000 378000000 317000
##  $ Knowledge     : chr  "136,3571429" "149,7142857" "136,3571429" "154,5" ...
##  $ YellowC       : int  5 3 2 4 1 2 2 1 3 0 ...
##  $ RedC          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FoulsC        : int  16 10 6 9 6 11 15 6 7 16 ...
##  $ FoulsT        : int  16 20 9 17 13 9 13 12 17 13 ...
##  $ TacklesW      : int  18 17 9 18 18 24 19 19 18 27 ...
##  $ TacklesAtt    : int  21 26 16 23 34 35 22 24 25 35 ...
##  $ TacklesWRatio : chr  "85,71%" "65,39%" "56,25%" "78,26%" ...
##  $ AirDuelW      : chr  "49%" "57%" "50%" "53%" ...
##  $ DribW         : int  16 11 11 9 22 9 18 17 12 13 ...
##  $ DribAtt       : int  33 25 23 22 35 20 30 23 21 21 ...
##  $ DribWRatio    : chr  "48,49%" "44,00%" "47,83%" "40,91%" ...
##  $ Group         : int  1 1 1 1 1 1 1 1 1 1 ...
```

We notice that percentages variables are stored as characters. we would like them to be represented as numerical [0,1]

```
df$BallPoss <- as.numeric(sub(",", ".", sub("%", "", df$BallPoss))) / 100
df$ShotsPrec <- as.numeric(sub(",", ".", sub("%", "", df$ShotsPrec))) / 100
df$PassPrec <- as.numeric(sub(",", ".", sub("%", "", df$PassPrec))) / 100
df$TacklesWRatio <- as.numeric(sub(",", ".", sub("%", "", df$TacklesWRatio))) / 100
```

```
df$AirDuelW <- as.numeric(sub(",", ".", sub("%", "", df$AirDuelW))) / 100
df$DribWRatio <- as.numeric(sub(",", ".", sub("%", "", df$DribWRatio))) / 100
```

the Knowledge column is a chr as well, let's handle it:

```
df$Knowledge <- as.numeric(sub(",", ".", df$Knowledge))
```

Columns ID and match are not useful for our model, let's remove them

```
df$ID <- NULL
df$Match <- NULL
```

We notice that the "result" column is redundant wrt goals made and taken, we remove it

```
df$Result <- NULL
```

Let's do some label encoding to handle cathegorical variables. We want in FieldAdvantage column, to show Home matches "H" as "1" and Away matches "A" as "0"

```
df$FieldAdvantage <- ifelse(df$FieldAdvantage == "H", 1,
                            ifelse(df$FieldAdvantage == "A", 0, NA))
```

We then want losses "L" as "-1", wins "W" as "1" and draws "D" as "0" in OutcomeWL column. This column can be redundant with respect to Points won. We will think about it later.

```
df$OutcomeWL <- ifelse(df$OutcomeWL == "W", 1,
                            ifelse(df$OutcomeWL == "D", 0,
                                   ifelse(df$OutcomeWL == "L", -1, NA)))
```

We confirm that column is now numerical

```
str(df)
```

```
## 'data.frame':    372 obs. of  26 variables:
##  $ FieldAdvantage: num  1 0 1 0 1 0 0 1 1 0 ...
##  $ OutcomeWL     : num  1 1 1 0 -1 -1 1 1 1 1 ...
##  $ PointsWon     : int  3 3 3 1 0 0 3 3 3 3 ...
##  $ GoalsMade     : int  2 2 3 2 1 0 2 3 4 2 ...
##  $ GoalsTaken    : int  1 1 0 2 2 3 0 0 0 0 ...
##  $ BallPoss      : num  0.617 0.624 0.601 0.551 0.527 0.502 0.591 0.549 0.439 0.452 ...
##  $ Shots         : int  16 13 22 28 12 9 22 16 14 13 ...
##  $ ShotsOnT      : int  7 4 10 7 4 2 9 6 6 7 ...
##  $ ShotsPrec     : num  0.438 0.308 0.455 0.25 0.333 ...
##  $ PassAtt       : int  480 488 545 416 474 445 460 458 364 325 ...
##  $ PassSucc      : int  556 560 626 488 540 535 554 553 447 409 ...
##  $ PassPrec      : num  0.863 0.871 0.871 0.853 0.878 ...
##  $ RosterQuality : int  365500000 378500000 365500000 387500000 422500000 383000000 378000000 317000
##  $ Knowledge     : num  136 150 136 154 128 ...
##  $ YellowC       : int  5 3 2 4 1 2 2 1 3 0 ...
##  $ RedC          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FoulsC        : int  16 10 6 9 6 11 15 6 7 16 ...
##  $ FoulsT        : int  16 20 9 17 13 9 13 12 17 13 ...
##  $ TacklesW      : int  18 17 9 18 18 24 19 19 18 27 ...
##  $ TacklesAtt    : int  21 26 16 23 34 35 22 24 25 35 ...
##  $ TacklesWRatio : num  0.857 0.654 0.562 0.783 0.529 ...
##  $ AirDuelW      : num  0.49 0.57 0.5 0.53 0.48 0.61 0.6 0.58 0.44 0.47 ...
##  $ DribW         : int  16 11 11 9 22 9 18 17 12 13 ...
##  $ DribAtt       : int  33 25 23 22 35 20 30 23 21 21 ...
##  $ DribWRatio    : num  0.485 0.44 0.478 0.409 0.629 ...
```

```
##  $ Group         : int  1 1 1 1 1 1 1 1 1 1 ...
```

Since our target variable is OutcomeWL, we check if it is balanced:

```
prop.table(table(df$OutcomeWL))
```

```
##
##        -1         0         1
## 0.2688172 0.1801075 0.5510753
```

The variable is not balanced, with a majority of winning matches (55%). The question that we want to answer by building a model with this target variable is: - Which are the most important factors that contribute to a win? We therefore decide to just concern ourselves with won matches a not-won matches We achieve this by joining draws and losses in a single category, thereby balancing the target variable.

```
df$OutcomeWL[df$OutcomeWL == -1] <- 0
```

We now have "0" for losses and draws and "1" for wins.

Step 3, Explore Data:

Let's see which features are correlated with each other:

S <- round(cov(df), 4) P <- round(cor(df), 4)

We use a heatmap to explore the correlations col<- colorRampPalette(c("blue", "white", "red"))(20) heatmap(x = P, col = col, symm = TRUE)

DS questions: - which are the most important factors that contribute to a win - do these factors change between group winners and control? - what can we learn from these statistics about the difference between control and winners in how they play the game? - Without using the roster quality feature is it possible to predict if a team is a winner or control? which are the features that contribute to this the most? - Is it possible to predict the amount of goals scored, or taken, by looking at these statistics? with which accuracy? - between the statistics that can be improved with training, which are the most relevant toward scoring more goals or winning a match?