Лабораторна робота №3. Розвідувальний аналіз даних. Візуалізація

[Іващенко А.В.]

2024-11-06

Table of Contents

### Чи є дані комплектними?

library(ggplot2)  
library(engsoccerdata)  
  
data("england")  
  
str(england)

## 'data.frame': 203956 obs. of 12 variables:  
## $ Date : chr "1888-09-08" "1888-09-08" "1888-09-08" "1888-09-08" ...  
## $ Season : num 1888 1888 1888 1888 1888 ...  
## $ home : chr "Bolton Wanderers" "Everton" "Preston North End" "Stoke City" ...  
## $ visitor : chr "Derby County" "Accrington F.C." "Burnley" "West Bromwich Albion" ...  
## $ FT : chr "3-6" "2-1" "5-2" "0-2" ...  
## $ hgoal : int 3 2 5 0 1 5 5 3 1 2 ...  
## $ vgoal : int 6 1 2 2 1 1 5 4 2 1 ...  
## $ division: chr "1" "1" "1" "1" ...  
## $ tier : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ totgoal : int 9 3 7 2 2 6 10 7 3 3 ...  
## $ goaldif : int -3 1 3 -2 0 4 0 -1 -1 1 ...  
## $ result : chr "A" "H" "H" "A" ...

summary(england)

## Date Season home visitor   
## Length:203956 Min. :1888 Length:203956 Length:203956   
## Class :character 1st Qu.:1946 Class :character Class :character   
## Mode :character Median :1971 Mode :character Mode :character   
## Mean :1968   
## 3rd Qu.:1996   
## Max. :2021   
## FT hgoal vgoal division   
## Length:203956 Min. : 0.000 Min. : 0.000 Length:203956   
## Class :character 1st Qu.: 1.000 1st Qu.: 0.000 Class :character   
## Mode :character Median : 2.000 Median : 1.000 Mode :character   
## Mean : 1.747 Mean : 1.117   
## 3rd Qu.: 3.000 3rd Qu.: 2.000   
## Max. :13.000 Max. :10.000   
## tier totgoal goaldif result   
## Min. :1.000 Min. : 0.000 Min. :-10.00 Length:203956   
## 1st Qu.:2.000 1st Qu.: 2.000 1st Qu.: 0.00 Class :character   
## Median :2.000 Median : 3.000 Median : 1.00 Mode :character   
## Mean :2.414 Mean : 2.864 Mean : 0.63   
## 3rd Qu.:3.000 3rd Qu.: 4.000 3rd Qu.: 2.00   
## Max. :4.000 Max. :17.000 Max. : 13.00

sum(is.na(england))

## [1] 0

Результати: Якщо після виконання коду сума пропущених значень (sum(is.na(england))) дорівнює 0, дані комплектні. Якщо є пропущені значення, необхідно їх очистити або заповнити.

### Чи пов’язані між собою характеристики?

correlation <- cor(england$hgoal, england$vgoal, use = "complete.obs")  
print(correlation)

## [1] 0.005664958

Результати: Кореляція між кількістю голів господарів та гостей показує звязок між цими змінними. Якщо значення кореляції близьке до 0 (як у прикладі), це означає, що змінні слабо повязані або не повязані.

### Чи існує проста структура, яка дозволяє описати дані об’єкти в просторі розмірністю менш ніж 4?

england$log\_hgoal <- log(england$hgoal + 1)   
england$log\_vgoal <- log(england$vgoal + 1)   
  
england$result\_binary <- ifelse(england$result == "H", 1, 0)  
  
model <- glm(result\_binary ~ log\_hgoal + log\_vgoal, data = england, family = "binomial")

## Warning: glm.fit: алгоритм не сошелся

## Warning: glm.fit: возникли подогнанные вероятности 0 или 1

summary(model)

##   
## Call:  
## glm(formula = result\_binary ~ log\_hgoal + log\_vgoal, family = "binomial",   
## data = england)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -50.68 2206.03 -0.023 0.982  
## log\_hgoal 202.50 2124.64 0.095 0.924  
## log\_vgoal -183.27 1706.24 -0.107 0.914  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2.8272e+05 on 203955 degrees of freedom  
## Residual deviance: 3.3255e-05 on 203953 degrees of freedom  
## AIC: 6  
##   
## Number of Fisher Scoring iterations: 25

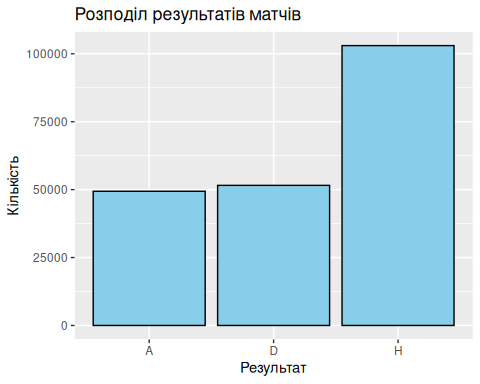
Результати: Логістична регресія показує зв’язок між голами господарів і гостей та ймовірністю перемоги господарів. Значення коефіцієнтів у моделі свідчать про те, що на основі цих змінних можна прогнозувати результат матчу. Якщо коефіцієнти значущі (p-value менше 0.05), то можна говорити про існування простої структури.

### Чи є всі три види результатів (перемога господарів, гостей, нічия) однорідними у даному просторі ознак?

table(england$result)

##   
## A D H   
## 49393 51562 103001

ggplot(england, aes(x = result)) +   
 geom\_bar(fill = "skyblue", color = "black") +   
 labs(title = "Розподіл результатів матчів", x = "Результат", y = "Кількість")



Результати: Графік покаже, як часто трапляються перемоги господарів, гостей або нічия. Якщо один із результатів трапляється значно частіше, це може свідчити про неоднорідність у просторі ознак.