Tidyverse Assignment

Alain T Kuiete

12/5/2019

How accurated is the FiveThirtyEigth Model on Predicting Soccer Match Scores

Library

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Loading the Datasets

```
soccer <- read.csv("https://raw.githubusercontent.com/AlainKuiete/DATA607/master/spi_matches.csv")
```

Introspecting the dataset

```
head(soccer)
```

```
##         date league_id                  league          team1
## 1 2016-08-12      1843          French Ligue 1         Bastia
## 2 2016-08-12      1843          French Ligue 1      AS Monaco
## 3 2016-08-13      2411 Barclays Premier League      Hull City
## 4 2016-08-13      2411 Barclays Premier League Crystal Palace
## 5 2016-08-13      2411 Barclays Premier League        Everton
## 6 2016-08-13      2411 Barclays Premier League  Middlesbrough
##                  team2  spi1  spi2  prob1  prob2 probtie proj_score1
## 1  Paris Saint-Germain 51.16 85.68 0.0463 0.8380  0.1157        0.91
## 2             Guingamp 68.85 56.48 0.5714 0.1669  0.2617        1.82
## 3       Leicester City 53.57 66.81 0.3459 0.3621  0.2921        1.16
## 4 West Bromwich Albion 55.19 58.66 0.4214 0.2939  0.2847        1.35
## 5    Tottenham Hotspur 68.02 73.25 0.3910 0.3401  0.2689        1.47
## 6           Stoke City 56.32 60.35 0.4380 0.2692  0.2927        1.30
##   proj_score2 importance1 importance2 score1 score2  xg1  xg2 nsxg1 nsxg2
## 1        2.36        32.4        67.7      0      1 0.97 0.63  0.43  0.45
## 2        0.86        53.7        22.9      2      2 2.45 0.77  1.75  0.42
## 3        1.24        38.1        22.2      2      1 0.85 2.77  0.17  1.25
## 4        1.14        43.6        34.6      0      1 1.11 0.68  0.84  1.60
## 5        1.38        31.9        48.0      1      1 0.73 1.11  0.88  1.81
## 6        1.01        33.9        32.5      1      1 1.40 0.55  1.13  1.06
##   adj_score1 adj_score2
## 1       0.00       1.05
## 2       2.10       2.10
## 3       2.10       1.05
## 4       0.00       1.05
## 5       1.05       1.05
## 6       1.05       1.05
```

```r
summary(soccer)
```

```
##          date         league_id                            league
##  2018-09-22:  160   Min.   :1818   English League Championship: 1666
##  2018-10-06:  151   1st Qu.:1849   Barclays Premier League    : 1520
##  2019-09-14:  148   Median :1871   French Ligue 1             : 1520
##  2019-09-21:  148   Mean   :2135   Italy Serie A              : 1520
##  2018-09-29:  145   3rd Qu.:2160   Spanish Primera Division   : 1520
##  2018-09-15:  144   Max.   :5641   Spanish Segunda Division   : 1398
##  (Other)   :31394                  (Other)                    :23146
##             team1                   team2            spi1
##  Arsenal        :   97   Arsenal        :   98   Min.   : 3.88
##  Atletico Madrid:   96   Atletico Madrid:   97   1st Qu.:31.18
##  Juventus       :   96   Real Madrid    :   96   Median :42.96
##  Real Madrid    :   96   Barcelona      :   95   Mean   :45.06
##  Barcelona      :   95   Juventus       :   95   3rd Qu.:58.57
##  Manchester City:   93   Liverpool      :   93   Max.   :96.57
##  (Other)        :31717   (Other)        :31716
##       spi2           prob1            prob2           probtie
##  Min.   : 4.04   Min.   :0.0271   Min.   :0.0032   Min.   :0.0000
##  1st Qu.:31.17   1st Qu.:0.3523   1st Qu.:0.2012   1st Qu.:0.2345
##  Median :42.86   Median :0.4439   Median :0.2785   Median :0.2610
##  Mean   :45.00   Mean   :0.4525   Mean   :0.2944   Mean   :0.2531
##  3rd Qu.:58.38   3rd Qu.:0.5417   3rd Qu.:0.3680   3rd Qu.:0.2824
##  Max.   :96.78   Max.   :0.9775   Max.   :0.8992   Max.   :0.4537
##
##   proj_score1     proj_score2     importance1      importance2
##  Min.   :0.250   Min.   :0.200   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:1.250   1st Qu.:0.890   1st Qu.: 10.90   1st Qu.: 10.50
##  Median :1.460   Median :1.110   Median : 26.20   Median : 25.30
##  Mean   :1.528   Mean   :1.156   Mean   : 31.29   Mean   : 30.58
##  3rd Qu.:1.730   3rd Qu.:1.370   3rd Qu.: 45.40   3rd Qu.: 44.50
##  Max.   :4.900   Max.   :4.010   Max.   :100.00   Max.   :100.00
##                                  NA's   :8768     NA's   :8768
##      score1          score2           xg1              xg2
##  Min.   : 0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 1.000   1st Qu.:0.000   1st Qu.:0.890   1st Qu.:0.610
##  Median : 1.000   Median :1.000   Median :1.380   Median :1.030
##  Mean   : 1.535   Mean   :1.171   Mean   :1.505   Mean   :1.155
##  3rd Qu.: 2.000   3rd Qu.:2.000   3rd Qu.:1.970   3rd Qu.:1.550
##  Max.   :10.000   Max.   :9.000   Max.   :7.070   Max.   :6.200  
##  NA's   :4526     NA's   :4526    NA's   :17070   NA's   :17070  
##      nsxg1           nsxg2          adj_score1      adj_score2   
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000  
##  1st Qu.:0.960   1st Qu.:0.730   1st Qu.:1.050   1st Qu.:0.000  
##  Median :1.320   Median :1.050   Median :1.050   Median :1.050  
##  Mean   :1.418   Mean   :1.131   Mean   :1.553   Mean   :1.179  
##  3rd Qu.:1.760   3rd Qu.:1.430   3rd Qu.:2.100   3rd Qu.:2.100  
##  Max.   :6.580   Max.   :5.920   Max.   :9.150   Max.   :7.930  
##  NA's   :17070   NA's   :17070   NA's   :17070   NA's   :17070
```

```
str(soccer)
```

```
## 'data.frame':    32290 obs. of  22 variables:
##  $ date       : Factor w/ 1191 levels "2016-08-12","2016-08-13",..: 1 1 2 2 2 2 2 2 2 2 ...
##  $ league_id  : int  1843 1843 2411 2411 2411 2411 2411 2411 1843 2411 ...
##  $ league     : Factor w/ 37 levels "Argentina Primera Division",..: 13 13 4 4 4 4 4 4 13 4 ...
##  $ team1      : Factor w/ 752 levels "1. FC Heidenheim 1846",..: 80 50 339 185 221 435 123 626 106 4
##  $ team2      : Factor w/ 752 levels "1. FC Heidenheim 1846",..: 506 313 393 736 682 650 660 733 638
##  $ spi1       : num  51.2 68.8 53.6 55.2 68 ...
##  $ spi2       : num  85.7 56.5 66.8 58.7 73.2 ...
##  $ prob1      : num  0.0463 0.5714 0.3459 0.4214 0.391 ...
##  $ prob2      : num  0.838 0.167 0.362 0.294 0.34 ...
##  $ probtie    : num  0.116 0.262 0.292 0.285 0.269 ...
##  $ proj_score1: num  0.91 1.82 1.16 1.35 1.47 1.3 1.37 1.91 1.39 2.69 ...
##  $ proj_score2: num  2.36 0.86 1.24 1.14 1.38 1.01 1.05 1.05 1.14 0.48 ...
##  $ importance1: num  32.4 53.7 38.1 43.6 31.9 33.9 36.5 34.1 37.9 73 ...
##  $ importance2: num  67.7 22.9 22.2 34.6 48 32.5 29.1 30.7 44.2 27 ...
##  $ score1     : int  0 2 2 0 1 1 0 1 3 2 ...
##  $ score2     : int  1 2 1 1 1 1 1 1 2 1 ...
##  $ xg1        : num  0.97 2.45 0.85 1.11 0.73 1.4 1.24 1.05 1.03 2.14 ...
##  $ xg2        : num  0.63 0.77 2.77 0.68 1.11 0.55 1.84 0.22 1.84 1.25 ...
##  $ nsxg1      : num  0.43 1.75 0.17 0.84 0.88 1.13 1.71 1.52 1.1 1.81 ...
##  $ nsxg2      : num  0.45 0.42 1.25 1.6 1.81 1.06 1.56 0.41 2.26 0.92 ...
##  $ adj_score1 : num  0 2.1 2.1 0 1.05 1.05 0 1.05 3.12 2.1 ...
##  $ adj_score2 : num  1.05 2.1 1.05 1.05 1.05 1.05 1.05 1.05 2.1 1.05 ...
```

Selecting Useful Variables Using Pipe and Predicting Score

```
pred.res <- soccer %>% select(spi1, spi2, score1, score2) %>%                          mutate(sc.p=if_else
head(pred.res)
```

```
##    spi1  spi2 score1 score2 sc.p
## 1 51.16 85.68      0      1    0
## 2 68.85 56.48      2      2    1
## 3 53.57 66.81      2      1    0
## 4 55.19 58.66      0      1    0
## 5 68.02 73.25      1      1    0
## 6 56.32 60.35      1      1    0
```

Actual scores

```
act.pred <- pred.res %>% mutate(sc.r=if_else(score1>=score2, 1, 0))%>% select(sc.r, sc.p)
head(act.pred)
```

```
##   sc.r sc.p
## 1    0    0
## 2    1    1
## 3    1    0
## 4    0    0
## 5    1    0
## 6    1    0
```

```
act.pred <- act.pred %>% mutate(diff=if_else(sc.r==sc.p, 1, 0))
```

```
res <- act.pred %>% group_by(diff)%>% summarise(count=n())
res
```

```
## # A tibble: 3 x 2
##    diff count
##   <dbl> <int>
## 1     0 10920
## 2     1 16844
## 3    NA  4526
```

Accuracy of prediction

```
pt <- as.numeric(filter(res, count, diff==1)[1,2])
pf <- as.numeric(filter(res, count, diff==0)[1,2])
```

```
pt
```

```
## [1] 16844
```

```
pf
```

```
## [1] 10920
```

```
ac <- pt/(pt+pf)
ac
```

```
## [1] 0.6066849
```

Overall, FiveThirtyEigth predicted at 60.6% the 2016 Europeean Leagues of Soccer.

Classification of teams by leagues in 2016 World Soccer leagues

```
s.teams <- select(soccer, league_id, league, team1, team2, score1, score2)
```

we can imbricate if else statements

```
s.teams <- s.teams %>% group_by(league_id, league, team1)%>% mutate(pt1=if_else(score1>score2,3, if_els
```

Points gained by each team

```
 s.teams <- s.teams %>% summarise_at( c("pt1","pt2"), sum, na.rm = TRUE) %>% mutate(pts = pt1+pt2)
```

Best teams by leagues in 2016 Wold Soccer leagues

```
s.teams %>% filter( pts==max(pts))
```

```
## # A tibble: 46 x 6
## # Groups:   league_id, league [37]
##    league_id league                     team1              pt1   pt2   pts
##        <int> <fct>                      <fct>            <dbl> <dbl> <dbl>
##  1      1818 UEFA Champions League      Juventus            40    16    56
##  2      1820 UEFA Europa League         Arsenal             39     9    48
##  3      1827 Austrian T-Mobile Bundesl~ SK Sturm Graz       70    52   122
##  4      1832 Belgian Jupiler League     Club Brugge         68    14    82
##  5      1832 Belgian Jupiler League     Standard Liege      65    17    82
##  6      1837 Danish SAS-Ligaen          Brondby             47    32    79
##  7      1837 Danish SAS-Ligaen          FC Copenhagen       71     8    79
##  8      1843 French Ligue 1             Paris Saint-Germ~  171    15   186
##  9      1844 French Ligue 2             Sochaux             73    61   134
## 10      1845 German Bundesliga          Bayern Munich      142    22   164
## # ... with 36 more rows
```