Predicting Turkish Super League Match Result Using Generalized Linear Model

Veli Kısa

23 Mayıs 2019

Introduction

This project aims to predict match score by using generalized linear model (GLM) in Turkish first division league.

Data contains 3-years match reasults (2015-2016,2016-2017, 2017-2018).

```
X2015_2016 <- read.csv("~/R/football prediction/turkishleague/2015-2016.csv")
X2016_2017 <- read.csv("~/R/football prediction/turkishleague/2016-2017.csv")
X2017_2018 <- read.csv("~/R/football prediction/turkishleague/2017-2018.csv")

result<-
rbind(X2015_2016[,c("HomeTeam","AwayTeam","FTHG","FTAG")],X2016_2017[c("HomeT
eam","AwayTeam","FTHG","FTAG")],X2017_2018[c("HomeTeam","AwayTeam","FTHG","FT
AG")])
result<-na.omit(result)
head(result)
```

```
##        HomeTeam             AwayTeam FTHG FTAG
## 1    Fenerbahce        Eskisehirspor    2    0
## 2    Buyuksehyr          Antalyaspor    2    3
## 3     Sivasspor          Galatasaray    2    2
## 4   Trabzonspor            Bursaspor    1    0
## 5 Gaziantepspor            Kasimpasa    0    3
## 6     Konyaspor Akhisar Belediyespor    1    1
```

Data frame is restricted to the columns in which we are interested. These columns,

HomeTeam= The team that is playing in the usual area that they play in AwayTeam= The team that is playing away from home. FTHG= Full Time Home Team Goals FTAG= Full Time Away Team Goals

```
mean(result[,3]) #Average of HomeGoals
```

```
## [1] 1.581699
```

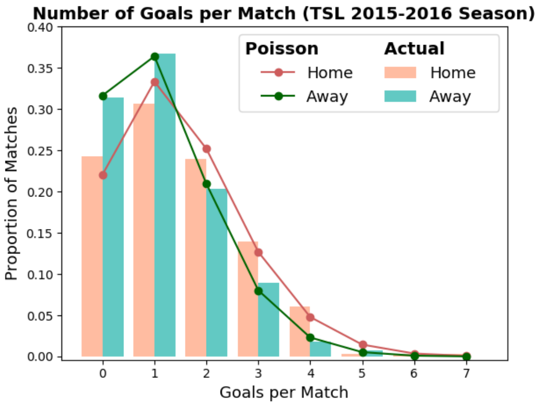
```
mean(result[,4]) #Average of AwayGoals
```

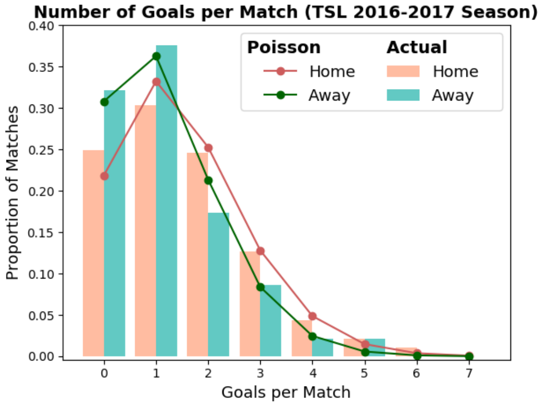
```
## [1] 1.208061
```

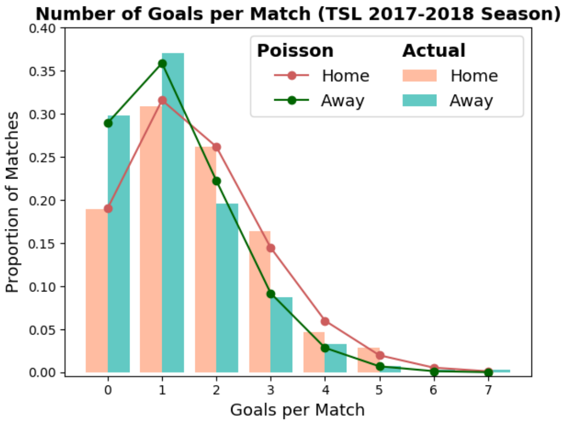
The home team scores more goals than away team. This can be descired as home field advantage. It’s a discrete probability distribution that describes the probability of the number of scores within a match (90 mins) with a known average rate of occurrence. We know that the goals don’t become more or less probable by the number of goals already scored in the match. So the number of goals is independent of time. Number of goals can be counted as function of average rate of goals.

$$f(x) = \frac{exp^{-\lambda}\lambda^{x}}{x!}, \quad \lambda > 0$$

λ represents the average number of goals in a match. “Home Goals” and “Away Goals” can be assumed as two independent Poisson distribution.







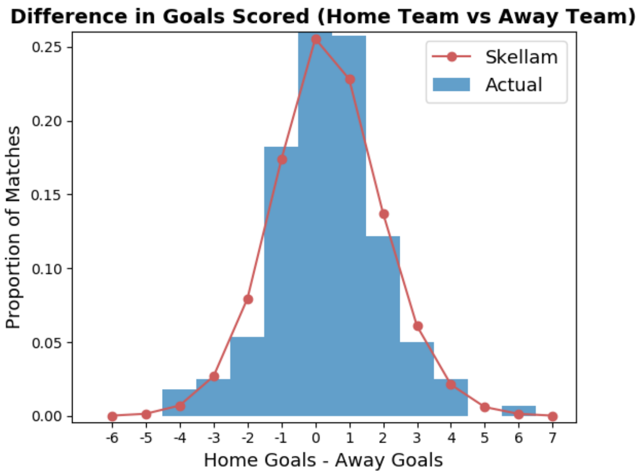
$P(HomeGoals > AwayGoals)$ means that home team wins the match. $P(AwayGoals > homeGoals)$ means that away team wins the match. $P(HomeGoals = AwayGoals)$ mean that draw.

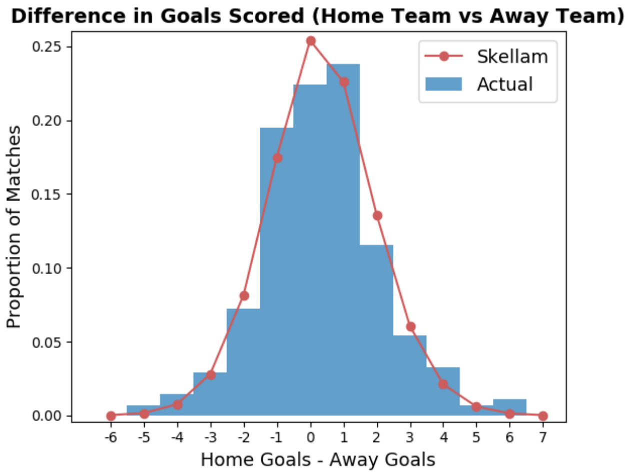
Assumption of the independence of goals scored by each teams,$P(A \cap B) = P(A)P(B)$ let us to construct the modal easily.

Skellam Distribution

The difference of two Poisson distributions called as Skellam distribution. Under the condition that difference of two Poisson distributions is 0,the draw status can be computed by skellam distribution.

```
#install.packages("skellam")
library(skellam)

## Warning: package 'skellam' was built under R version 3.4.4

dskellam(0,lambda1 = mean(result[,3]),lambda2 = mean(result[,4]))
#probability of draw

## [1] 0.2483468

dskellam(1,lambda1 = mean(result[,3]),lambda2 = mean(result[,4]))
#probability of home team winning by 1 goal

## [1] 0.2247654
```





Building a Model

```
colnames(result)[3:4]<-c("HomeGoals","AwayGoals")
model_data1<-result[,c(1,2,3)]
model_data1$home<-1
colnames(model_data1)[1:3]<-c("team","opponent","goals")
model_data2<-result[,c(2,1,4)]
model_data2$home<-0
colnames(model_data2)[1:3]<-c("team","opponent","goals")
model_data<-rbind(model_data1,model_data2)

p_model<-glm(formula = goals~home+team+opponent,family = poisson,data =
model_data)

summary(p_model)

## 
## Call:
## glm(formula = goals ~ home + team + opponent, family = poisson, 
##     data = model_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2056  -1.1904  -0.1018   0.5152   3.2042  
## 
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                 0.06359    0.12600   0.505 0.613804    
## home                        0.26948    0.03988   6.757 1.41e-11 ***
## teamAntalyaspor             0.06526    0.12152   0.537 0.591267    
## teamBesiktas                0.48171    0.11056   4.357 1.32e-05 ***
## teamBursaspor              -0.05237    0.12526  -0.418 0.675885    
## teamBuyuksehyr              0.28982    0.11490   2.522 0.011658 *  
## teamEskisehirspor          -0.10469    0.18272  -0.573 0.566687    
## teamFenerbahce              0.38977    0.11254   3.463 0.000534 ***
## teamGalatasaray             0.45541    0.11138   4.089 4.34e-05 ***
## teamGaziantepspor          -0.36203    0.15506  -2.335 0.019555 *  
## teamGenclerbirligi         -0.16721    0.12864  -1.300 0.193670    
## teamKasimpasa               0.15272    0.11900   1.283 0.199359    
## teamKayserispor            -0.12218    0.12746  -0.959 0.337775    
## teamKonyaspor              -0.08556    0.12576  -0.680 0.496302    
## teamMersin Idman Yurdu     -0.32624    0.20001  -1.631 0.102861    
## teamOsmanlispor             0.04666    0.12194   0.383 0.701974    
## teamRizespor               -0.06183    0.14034  -0.441 0.659507    
## teamSivasspor              -0.09575    0.14248  -0.672 0.501542    
## teamTrabzonspor             0.07658    0.12110   0.632 0.527163    
## teamAdanaspor              -0.31085    0.19732  -1.575 0.115182    
## teamAlanyaspor              0.23944    0.12975   1.845 0.064988 .  
## teamKarabukspor            -0.38776    0.15781  -2.457 0.014002 *  
## teamGoztep                  0.13326    0.16774   0.794 0.426933    
## teamYeni Malatyaspor       -0.12796    0.18451  -0.694 0.487996    
## opponentAntalyaspor         0.10829    0.11841   0.915 0.360445    
## opponentBesiktas           -0.32461    0.13394  -2.424 0.015369 *  
## opponentBursaspor           0.16598    0.11665   1.423 0.154769    
## opponentBuyuksehyr         -0.30927    0.13270  -2.331 0.019777 *  
## opponentEskisehirspor       0.35513    0.15196   2.337 0.019439 *  
## opponentFenerbahce         -0.33256    0.13392  -2.483 0.013019 *  
## opponentGalatasaray        -0.07682    0.12494  -0.615 0.538632    
## opponentGaziantepspor       0.23280    0.12691   1.834 0.066607 .  
## opponentGenclerbirligi     -0.05344    0.12283  -0.435 0.663473    
## opponentKasimpasa           0.08677    0.11918   0.728 0.466566    
## opponentKayserispor         0.11807    0.11785   1.002 0.316389    
## opponentKonyaspor           -0.12960    0.12541  -1.033 0.301413    
## opponentMersin Idman Yurdu   0.44916    0.14681   3.059 0.002218 ** 
## opponentOsmanlispor          0.03870    0.12038   0.321 0.747835    
## opponentRizespor             0.11625    0.13159   0.883 0.377028    
## opponentSivasspor            0.09929    0.13155   0.755 0.450373    
## opponentTrabzonspor          0.06149    0.11977   0.513 0.607673    
## opponentAdanaspor            0.30605    0.15351   1.994 0.046181 *  
## opponentAlanyaspor           0.31998    0.12443   2.572 0.010122 *  
## opponentKarabukspor          0.36570    0.12195   2.999 0.002710 ** 
## opponentGoztep               0.08276    0.16572   0.499 0.617473    
## opponentYeni Malatyaspor    -0.03654    0.17230  -0.212 0.832042    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 2322.6  on 1835  degrees of freedom
## Residual deviance: 2014.7  on 1790  degrees of freedom
## AIC: 5406.5
## 
## Number of Fisher Scoring iterations: 5
```

The result of matches is modeled by glm with poisson family and log link function. When we examine the coefficients table, there are both positive and negative values. Similar to logistic regression, we take the exponent of the parameter values. A positive value implies more goals, while negative value that close to zero represent neutral effect. According to coefficents table, “home” has coefficient of 0.26948 and we can say that home teams generally score more goals with $e^{0.26946} = 1.34913$ times more likely than the away teams. But this is not the same for all teams. For example, Besiktas and Fenerbahce are better scorers than average with 0.79256 and 0.70061 parameters respectively, while Kasimpasa is worse scorer than average with -0.21928 parameter. Eventually, the *opponent* values penalize or reward teams based on the quality of the opposition.

Predictions

We now start making some predictions for the upcoming matches. Firstly,let us compute what is the expected average number of goals based on poisson distribution by selected teams.

```
predict.glm(p_model,data.frame(team="Besiktas",opponent="Bursaspor",home=1),t
ype ="response")[1]
```

```
##        1 
## 2.666494
```

```
predict.glm(p_model,data.frame(team="Bursaspor",opponent="Besiktas",home=0),t
ype ="response")[1]
```

```
##         1 
## 0.7309635
```

We have two Poisson distributions. By creating a function called “sim_match” we can calculate the probability of various events.

```
sim_match<- function(f_model,homeT,awayT,max_goals){
  homegoals_ave<-
predict.glm(p_model,data.frame(team=homeT,opponent=awayT,home=1),type
="response")[1]
  awaygoals_ave<-
predict.glm(p_model,data.frame(team=awayT,opponent=homeT,home=0),type
="response")[1]
  team_pred<-matrix(NA,max_goals+1,2)
  for (i in 0:max_goals){
    team_pred[i+1,1]<-dpois(i,homegoals_ave)
    team_pred[i+1,2]<-dpois(i,awaygoals_ave)
    
  }
  match_result_matrix<-outer(team_pred[,1],team_pred[,2],FUN = "*")
  return(list(team_pred,match_result_matrix))
  
}
sim_match(f_model = p_model,homeT = "Besiktas",awayT =
"Kasimpasa",max_goals=3)
```

```
## [[1]]
##            [,1]       [,2]
## [1,] 0.08514221 0.40764584
## [2,] 0.20974207 0.36580365
## [3,] 0.25834271 0.16412814
## [4,] 0.21213660 0.04909382
## 
## [[2]]
##            [,1]       [,2]       [,3]        [,4]
## [1,] 0.03470787 0.03114533 0.01397423 0.004179956
## [2,] 0.08550049 0.07672442 0.03442458 0.010297040
## [3,] 0.10531233 0.09450271 0.04240131 0.012683030
## [4,] 0.08647660 0.07760034 0.03481759 0.010414596
```

The first column of the matrix 1 represents the probability of Besiktas (HomeTeam) scoring a specific number of goals according to rows, while the second column of the matrix 1 indicates the Kasimpasa (AwayTeam) scoring. From two independent poisson distribution, we create square matrix 2 that shows the match result score by multiplying two vector each other.

According to matrix 2, rows represent the Besiktas (HomeTeam) and columns indicate the Kasimpasa (AwayTeam). When we analyze matrix 2 in depth , along the diagonal both teams score the same number of goals. For example, the probability of draw without scores is $P(0-0)=0.0347$, while the odds of a draw with single goal is $P(1-1)=0.0767$. So we can calculate the odds of draw by summing all the diagonal values.

The area under the diagonal shows us the situations that Besiktas won, while the area above the diagonal shows us the situations Kasimpasa won.

It can also be calculated over 2.5 goals or below from the matrix 2.

Now let us let us calculate the probability of Besiktas (HomeTeam) winning the match under the maximum of 10 goals,

```
bes_kas<-sim_match(f_model = p_model,homeT = "Besiktas",awayT =
"Kasimpasa",max_goals=10)
bes_win.mat<-do.call(rbind,lapply(bes_kas[2],matrix,ncol=11,byrow=FALSE))
sum(bes_win.mat[lower.tri(bes_win.mat)])

## [1] 0.7168926
```

And the probability of Kasimpasa winning the match under the maximum of 10 goals,

```
kas_win.mat<-do.call(rbind,lapply(bes_kas[2],matrix,ncol=11,byrow=FALSE))
sum(kas_win.mat[upper.tri(kas_win.mat)])

## [1] 0.1172307
```

Last, the probability of draw,

```
draw.mat<-do.call(rbind,lapply(bes_kas[2],matrix,ncol=11,byrow=FALSE))
sum(diag(draw.mat))

## [1] 0.1658225
```