# Games Code

Rafael Guimaraes, Yilang Tang and Yuanxi Yao

3/28/2017

## Question 1

Write down the equation to predict thw chnce of winning.

```
library(nnet)
df.train$prog2<-relevel(df.train$Match_O, ref="Loss")
train<-
multinom(prog2~HTGD+RED.H+RED.A+POINTS_H+POINTS_A+TOTAL_H_P+TOTAL_A_P+FGS.0+F
GS.1,
               data=df.train)
```

```
## # weights:  33 (20 variable)
## initial  value 1669.890679 
## iter  10 value 1350.231305
## iter  20 value 1141.100550
## iter  30 value 1133.951679
## final  value 1133.951565 
## converged
```

```
z <- summary(train)$coefficients/summary(train)$standard.errors
p <- (1 - pnorm(abs(z), 0, 1)) * 2
```

Now to read the result:

```
summary(train)
```

```
## Call:
## multinom(formula = prog2 ~ HTGD + RED.H + RED.A + POINTS_H + 
##     POINTS_A + TOTAL_H_P + TOTAL_A_P + FGS.0 + FGS.1, data = df.train)
## 
## Coefficients:
##      (Intercept)      HTGD      RED.H     RED.A   POINTS_H    POINTS_A
## Draw    3.534989 0.5111854  0.3005473 0.4629235 0.02403643 -0.01779833
## Win     3.313188 1.6183962 -0.8102867 0.9832042 0.03520821 -0.03451276
##          TOTAL_H_P   TOTAL_A_P     FGS.0     FGS.1
## Draw -0.0001797641 -0.01023752 -3.520911 -2.819490
## Win   0.0102033252 -0.01468249 -3.319533 -2.473055
## 
## Std. Errors:
##      (Intercept)      HTGD     RED.H     RED.A    POINTS_H    POINTS_A
## Draw   0.4620808 0.1201270 0.5722288 0.5396210 0.008885585 0.007916140
## Win    0.4703446 0.1434815 0.7431552 0.5465152 0.009217549 0.008585501
##        TOTAL_H_P   TOTAL_A_P     FGS.0     FGS.1
## Draw 0.003570172 0.003823573 0.409595 0.4261340
## Win  0.003795461 0.003954150 0.413159 0.4299782
##
## Residual Deviance: 2267.903
## AIC: 2307.903
```

## Question 2:

According to your model, how do red cards conceded by the Home and Away team influence the outcome of a match. Your response should not be numeric, but qualitative. Based on your understanding of the sport, speculate about possible reasons for these findings.

**Answer:** Influence the outcome of a match. Your response should not be numeric, but qualitative. Based on your understanding of the sport, speculate about possible reasons for these findings. A negative coefficient (-0.81) between red cards conceded by Home team and winning probability implies that the number of red cards negatively related to the winning probability of Home team . it means as the number of red card increases, the winning probability of Home team goes down. A positive coefficient (0.983) be between red cards conceded by Away team and winning probability shows that the number of red cards positively related to the winning probability of Away team, as the number of red card increases, the winning probability of Away team also increases.

positive coefficient between red cards and the draw probability of both teams shows that the more the red cards conceded, the higher chance of drawing for both teams.

## Question 3:

Under what circumstances is it informative to include points scored by a team in the prior season to predict the outcome of a match?

**Answer:** According to result of Z test, we got to know that a p value of 0.9598 (p>0.05)in the draw model for Total_H_P means that prior points scored by Home team are not relevant to the draw probability of Home team and a p value of 0.007 (p<0.05)in the win model for Total_H_P are significantly relevant to the winning probability of Home team. While prior points scored by Away team does highly relate to the draw and winning probability(p draw<0.05, P win<0.05) same theory for the relation between the points_H , points_A and the draw an win probability. For Home team, the points earned in the league highly relevant to the draw and winning probability while for Away team, , the points earned in the league does not relate to the draw and winning probability at all. So when we predict the winning chance of Home team, using the data of points earned in the league and total prior scores would be both informative; while for Away team, it is more informative to use the total prior scores to predict the outcome of the match instead of using points earning in the league.

## Question 4

Using the multinomial logit model, compute the probability of a home team win for a match with these attributes:

Because of the p-value we removed the red cars variables. Below you can see our multinomial logit model.

```
q4<-multinom(prog2~HTGD+POINTS_H+POINTS_A+TOTAL_H_P+TOTAL_A_P+FGS.1,
            data=df.train)

## # weights:  24 (14 variable)
## initial  value 1669.890679
## iter  10 value 1364.953045
## iter  20 value 1218.298245
## final  value 1218.295583
## converged

summary(q4)

## Call:
## multinom(formula = prog2 ~ HTGD + POINTS_H + POINTS_A + TOTAL_H_P +
##     TOTAL_A_P + FGS.1, data = df.train)
##
## Coefficients:
##      (Intercept)     HTGD   POINTS_H    POINTS_A   TOTAL_H_P   TOTAL_A_P
## Draw   0.8548861 1.081792 0.02547116 -0.01860787 0.001996219 -0.01024537
## Win    0.7437707 2.115211 0.03613106 -0.03455355 0.012217068 -0.01494357
##           FGS.1
## Draw -0.4988693
## Win  -0.2132995
##
## Std. Errors:
##      (Intercept)      HTGD    POINTS_H    POINTS_A   TOTAL_H_P   TOTAL_A_P
## Draw   0.2501754 0.1194599 0.008476699 0.007471786 0.003349770 0.003648814
## Win    0.2652986 0.1467775 0.008924824 0.008292581 0.003639906 0.003803931
##          FGS.1
## Draw 0.2255361
## Win  0.2288930
##
## Residual Deviance: 2436.591
## AIC: 2464.591
```

Calculate the chance of winning by weithing them:

```
Pweight <-
coef(q4)[2,2]*3+coef(q4)[2,4]*18+coef(q4)[2,5]*15+coef(q4)[2,6]*39+coef(q4)[2
,2]*32

Pweight
```

```
## [1] 73.01088
```

## Question 5

If the first goal is scored by the away team, is it advisable to bet in favor of the away team? answer by controlling for all the other variables in the regression model.

**Answer:** It is not advisable to bet in favor of the away team if the first goal is scored by the away team. Because the model shows the first goal score is not dependent from the winning of the team (p>0.05).

## Question 6

What conclusion can you derive from the classfication tables shown in Exhibit 8? Is it advisable to be on draws?(based on the model developed)?

**Answer:** His classification tables shown in Exhibit 8 implies that the model did the best in predicting the loss (79.8%)and also had high accuracy (78.9%)in predicting the winning., but it did not perform well in draw(only 25.5%). So it is not advisable to be on draw since it had low correctness of prediction.

## Question 7

Using the Decision Tree from our R Script, write out several practical rules for betting. Be selective; if there is a tree branch that essentially says “don’t bet at all”, then omit that rule.

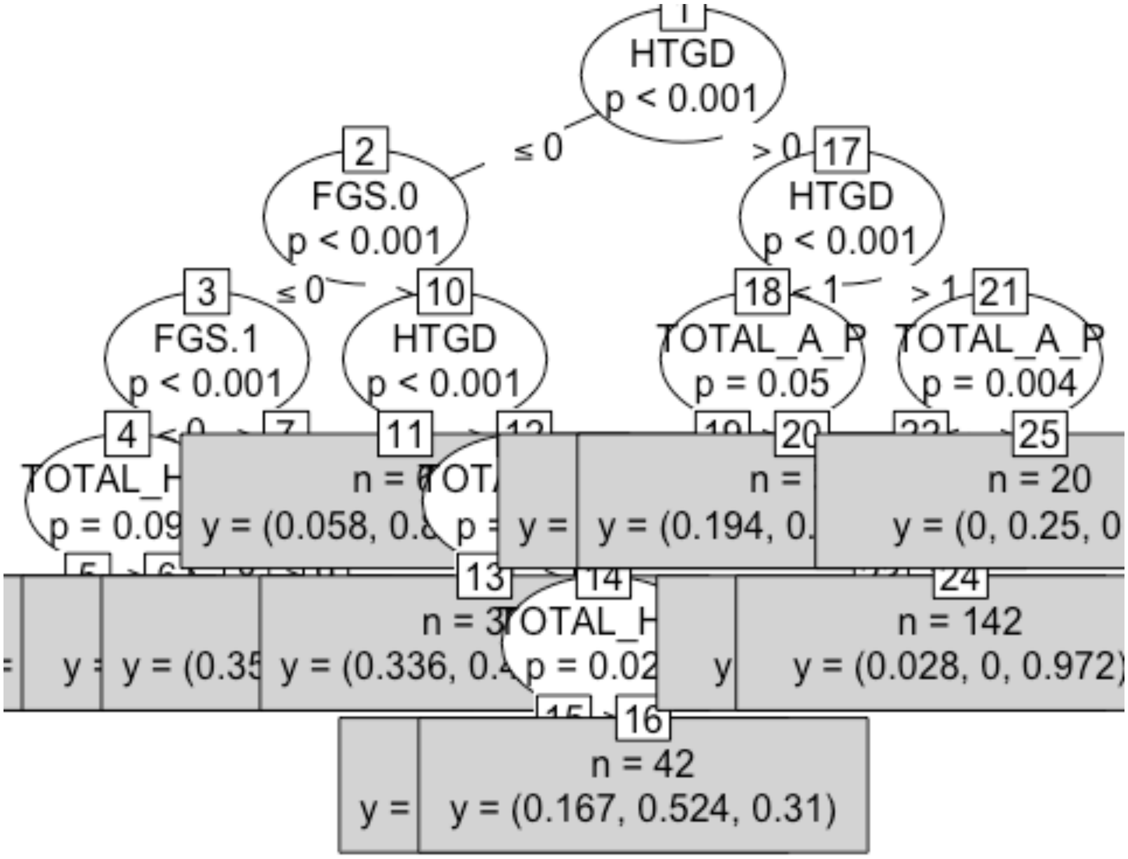
**Answer:**

Based on the tree plot below: 1. Betting on lose when HTGD is greater than one and at the same time TOTAL_A_P is greater than 64, you’ll have a probability of greater than 75% to win the bet. 2. Betting on lose when HTGD is greater than zero but lower than one and at the same time TOTAL_A_P is smaller than 82, you’ll have a probability of greater than 79.5% to win the bet. 3. Betting on win when HTGD is lower or equal to -2 and at the same time FGS.0 is greater than 0, you’ll have a probability of greater than 89.9% to win the bet

```
## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

## 
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
##
##   Conditional inference tree with 13 terminal nodes
##
## Response:  Match_O
## Inputs:  HTGD, RED.H, RED.A, POINTS_H, POINTS_A, TOTAL_H_P, TOTAL_A_P,
FGS.0, FGS.1
## Number of observations:  1520
##
## 1) HTGD <= 0; criterion = 1, statistic = 512.524
##   2) FGS.0 <= 0; criterion = 1, statistic = 254.288
##     3) FGS.1 <= 0; criterion = 1, statistic = 76.495
##       4) TOTAL_H_P <= 76; criterion = 0.908, statistic = 9.072
##         5)*  weights = 269
##       4) TOTAL_H_P > 76
##         6)*  weights = 45
##     3) FGS.1 > 0
##       7) TOTAL_A_P <= 61; criterion = 0.993, statistic = 14.174
##         8)*  weights = 70
##       7) TOTAL_A_P > 61
##         9)*  weights = 37
##   2) FGS.0 > 0
##     10) HTGD <= -2; criterion = 1, statistic = 31.087
##       11)*  weights = 69
##     10) HTGD > -2
##       12) TOTAL_A_P <= 64; criterion = 0.95, statistic = 10.335
##         13)*  weights = 327
##       12) TOTAL_A_P > 64
##         14) TOTAL_H_P <= 56; criterion = 0.974, statistic = 11.695
##           15)*  weights = 110
##         14) TOTAL_H_P > 56
##           16)*  weights = 42
## 1) HTGD > 0
##   17) HTGD <= 1; criterion = 0.999, statistic = 19.558
##     18) TOTAL_A_P <= 82; criterion = 0.95, statistic = 10.339
##       19)*  weights = 347
##     18) TOTAL_A_P > 82
##       20)*  weights = 31
##   17) HTGD > 1
##     21) TOTAL_A_P <= 64; criterion = 0.996, statistic = 15.34
##       22) TOTAL_H_P <= 0; criterion = 0.988, statistic = 10.269
##         23)*  weights = 11
##       22) TOTAL_H_P > 0
##         24)*  weights = 142
##       21) TOTAL_A_P > 64
##         25)*  weights = 20
```



## Question 8

For these, use the R output instead of the Exhibits mentioned in the case.Exibits 10 lists 20 matches played over two weekends in 2012 along with the values of the covariates. Use multinomial logistic regression to predict the match outcome in all 20 cases listed in exhibit 10

**Answer:**

```
df.test$prog2<-relevel(df.test$MATCH_O, ref="Loss")
Win <- multinom(prog2~HTGD+POINTS_H+POINTS_A+TOTAL_A_P+TOTAL_H_P+FGS.0+FGS.1,
data=df.test)
```

```
## # weights:  27 (16 variable)
## initial  value 21.972246
## iter  10 value 13.012464
## iter  20 value 0.191895
## iter  30 value 0.000232
## final  value 0.000051
## converged
```

```
pred<-predict(Win, df.test)
pred

##  [1] Loss Win  Win  Win  Draw Loss Loss Win  Draw Win  Loss Win  Loss Loss
## [15] Draw Win  Win  Loss Draw Loss
## Levels: Loss Draw Win
```

## Question 9

Apply the CHAID decision tree on the 20 matches listed in exhibit 10 and compare the results with your answers obtained using multinomial logistic regression.

**Answer:**

```
df.test$RED.A <- df.test$RED_A
df.test$RED.H <- df.test$RED_H
datactree2 <-
ctree(Match_O~HTGD+RED.H+RED.A+POINTS_H+POINTS_A+TOTAL_H_P+TOTAL_A_P+FGS.0+FG
S.1,
                   df.train, controls=ctree_control(mincriterion=0.9,
minsplit=50))
pred<-predict(datactree2, df.test)
```

Showing the results of the prediction:

```
pred

##  [1] Loss Draw Loss Draw Loss Loss Draw Win  Win  Loss Loss Win  Loss Loss
## [15] Draw Win  Draw Loss Draw Loss
## Levels: Draw Loss Win
```

## Question 10

if Peter were to choose one match from the list of 20 matches for betting, which match should he choose? Discuss the reasons for your suggestion.

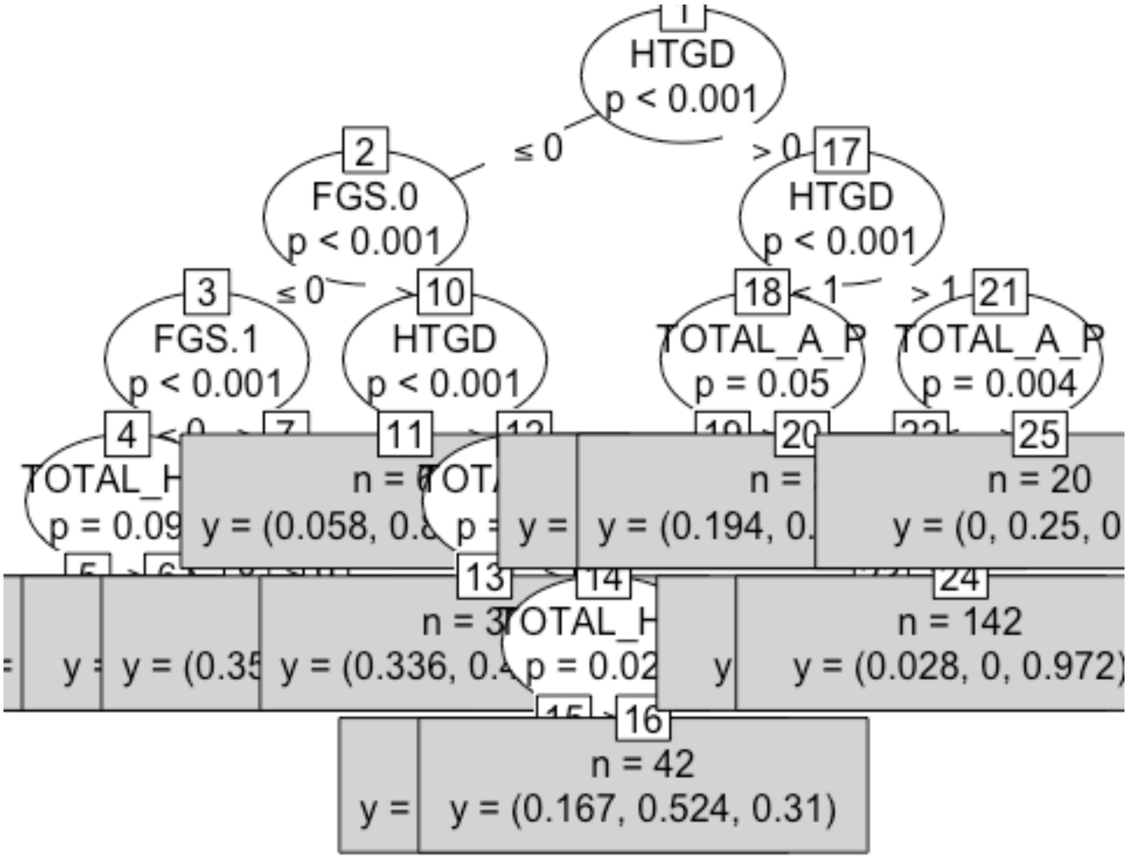
**Answer:** Everton vs. Southampton City (12)

Peter should choose No.12 match(Everton vs. Southampton). Because it has the highest probability of winning. According to the tree, the rout starts from 1 to 24 has the highest probability to win and it was supposed to have positive HTGD, Total_A_P lower than 64 and TOTAL_H_P higher than 0. So any match matches this factors will have the highest probability to win. 12 has 2 HTGD, 0 TOTAL_A_P, 56 total_H_P, which is the best choice.

```
print(datactree2, newdata = df.test)

##
##   Conditional inference tree with 13 terminal nodes
##
## Response:  Match_O
## Inputs:  HTGD, RED.H, RED.A, POINTS_H, POINTS_A, TOTAL_H_P, TOTAL_A_P,
FGS.0, FGS.1
```

```
## Number of observations:  1520
##
## 1) HTGD <= 0; criterion = 1, statistic = 512.524
##   2) FGS.0 <= 0; criterion = 1, statistic = 254.288
##     3) FGS.1 <= 0; criterion = 1, statistic = 76.495
##       4) TOTAL_H_P <= 76; criterion = 0.908, statistic = 9.072
##         5)*  weights = 269
##       4) TOTAL_H_P > 76
##         6)*  weights = 45
##     3) FGS.1 > 0
##       7) TOTAL_A_P <= 61; criterion = 0.993, statistic = 14.174
##         8)*  weights = 70
##       7) TOTAL_A_P > 61
##         9)*  weights = 37
##   2) FGS.0 > 0
##     10) HTGD <= -2; criterion = 1, statistic = 31.087
##       11)*  weights = 69
##     10) HTGD > -2
##       12) TOTAL_A_P <= 64; criterion = 0.95, statistic = 10.335
##         13)*  weights = 327
##       12) TOTAL_A_P > 64
##         14) TOTAL_H_P <= 56; criterion = 0.974, statistic = 11.695
##           15)*  weights = 110
##         14) TOTAL_H_P > 56
##           16)*  weights = 42
## 1) HTGD > 0
##   17) HTGD <= 1; criterion = 0.999, statistic = 19.558
##     18) TOTAL_A_P <= 82; criterion = 0.95, statistic = 10.339
##       19)*  weights = 347
##     18) TOTAL_A_P > 82
##       20)*  weights = 31
##   17) HTGD > 1
##     21) TOTAL_A_P <= 64; criterion = 0.996, statistic = 15.34
##       22) TOTAL_H_P <= 0; criterion = 0.988, statistic = 10.269
##         23)*  weights = 11
##       22) TOTAL_H_P > 0
##         24)*  weights = 142
##     21) TOTAL_A_P > 64
##       25)*  weights = 20
```

```
plot(datactree2,type="simple")
```


1
HTGD
p < 0.001
≤ 0
> 0
2
FGS.0
p < 0.001
17
HTGD
p < 0.001
3
≤ 0
10
18
≤ 1
> 1
21
FGS.1
p < 0.001
HTGD
p < 0.001
TOTAL_A_P
p = 0.05
TOTAL_A_P
p = 0.004
4
≤ 0
7
11
12
19
20
22
25
TOTAL_H
p = 0.09
n = 6
y = (0.058, 0.8
TOTA
p =
y =
n =
y = (0.194, 0.
n = 20
y = (0, 0.25, 0
5
6
13
14
24
y =
y = (0.35
n = 3
y = (0.336, 0.4
TOTAL_H
p = 0.02
y
n = 142
y = (0.028, 0, 0.972)
15
16
n = 42
y =
y = (0.167, 0.524, 0.31)