**Team Members**

Chintala Dorababu

Sai Hruthik Reddy Varakantham

Abigna Gutta

**Assignment - 4**

**Question 1**

Given that we need to find the equation to find the probability of win for the home team Peter can use given that alpha = 0.05

We know that P(Y=1) = bo + (𝚺 bj \* xj) where j = 1 to k

We exclude variables which have Sig value(p-value >0.05) , because these values establish that we fail to reject the null hypothesis(Ho) for Alternate Hypothesis(Ha) indicating that there is no relationship between the predictor and target variable.

We exclude variables RED\_H, RED\_A since they have values p values 0.275, 0.072 which are greater than Alpha value of 0.05

Therefore we have the equation for probability of home team win as:

P(Home team winning/Away Team winning) = 3.313 + (0.035) \* POINTS\_H + (-0.035) \* POINTS\_A + (1.618) \* HTGD + (-0.015) \* TOTAL\_A\_P + (-2.473) \* [FGS=1] + (-3.320) \* [FGS=0] + (0.010) \* TOTAL\_H\_P

**Question 2**

* We take an R-markdown before looking at values in **Exhibit 7** to explain how less impact a red card would have on the home team or the away team winning the game.

## **R Markdown**

*#Question 2*

setwd("~/Desktop/Data\_Mining/R-Files")

pacman::p\_load("pacman","rio","tidyverse","readxl","rpart","rpart.plot")

Homework\_4 <- read\_xlsx("Assignment\_4.xlsx")

table(Homework\_4$`RED-H` ,Homework\_4$Match\_O)

##

## 0 1 2

## 0 383 395 719

## 1 9 9 5

table(Homework\_4$`RED-A` ,Homework\_4$Match\_O)

##

## 0 1 2

## 0 385 393 702

## 1 7 11 21

## 2 0 0 1

* Match Outcome(Vertical) RED\_H(Home) RED\_A(Away)

0 -> Lose 0 -> no red card 0 -> No red card

1 -> Draw 1 -> 1 red card 1 -> 1 red card

2 -> Win 2 -> 2 red cards

* The following matrix: the horizontal represents the match outcome and the vertical line represents the number of red cards given to the home team and the away team.
* So in the first table we can see that there were only 9 cases out of them where a red card/s has made a difference in the home team losing and 7 cases which resulted in Away team losing because there was a red card given to them.
* In **Exhibit 7** The significance of RED\_H about 0.599,0.275 when Match\_O = 1,2 and RED\_A about 0.391 and 0.072 when Match\_O = 1,2 shows they are not statistically significant while considering them.

**Question 3**

* The points in previous matches need to be considered since the p value in TOTAL\_H\_P and TOTAL\_A\_P have p values less than Alpha.
* Except for the case when TOTAL\_H\_P with Match\_O = 1 we have Sig.(p value) of 0.960. We can consider this as an exception.
* All other values are as follows:

| Match\_O = 1 | TOTAL\_A\_P Sig. = 0.007 < α=0.05 |
| --- | --- |
| Match\_O=2 | TOTAL\_H\_P Sig. = 0.007 < α=0.05 |
| Match\_O=2 | TOTAL\_H\_P Sig. = 0.007 < α=0.05 |

* Therefore these are all statistically significant values and they need to be considered.

**Question 4**

* We also considered the draw probability since there can be only two values of input in a categorical variable therefore

T**he equation for the probability that the home team will win the match would be**

P = 3.13 + HTGD \* 2 + FGS \* 1 + 0 \* RED\_H + 0 \* RED\_A + 15 \* POINTS\_H + POINTS\_A \* 18 + 40 \* TOTAL\_H\_P + 30 \* TOTAL\_A\_P

P(Win/Loss) = 3.13 + 2 \* 1.618 + (-2.473) \* 1 +0.035 \* 15 + 40 \* 0.010 + (-0.015) \* 30 + (-.0.035) \* 18 = 3.738

P(Draw/Loss) = 3.13 + 2 \* 0.511 + (-2.819) \* 1 + 15 \* 0.024 + (-0.018) \* 18 +(0)\* 40 + (-0.010) \* 30 = 1.069

Total Probability = e^(3.738) /(1+e^(3.738) +e^(1.069)) = 91.48 % chance of home team winning

**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.

From **Exhibit 7** we can see that when away team scores a goal the [FGS = 0] and the match outcome we consider is a win (Match\_O = 2)

The formula would e^(1\* [FGS=0]) = e^(1\*-3.32) = 0.036

which is a 3.6% increase of the away team winning

* It is not advisable to bet in favor of the away team due to the odds of them winning increases by 3.6 %

**Question 6**

There are about 103 draws where the predicted value is a draw and also the observed value is a draw therefore we consider the correct percentage which is the assurance that we get to see a draw in the end which is 25.5 %

Here we get to that conclusion by

Draw Correct percentage = 103/(103+120+181) = 25.495 % ~ 25.5 %

**Question 7**

As you see the very first split decides on the yes or no and with the chi-square test, we can see that the p-value is indeed very small (0.0)

And the algorithm has divided the most predictive way to divide our data into different buckets.

Below is some analysis our group has done:

* Node 5: HTGD is -2, -3, or -5, the best bet would be on the away team with a percentage of 91.3.
* Node 6: If the HTGD is 0 and the FGS is 2, then bet on a draw because the draw percentage is about 49.7 percent and the win percentage is about 48.1% with no goals scored.
* Node 7: If the FGS and HTGD are 0 is 1 that is of home team scores, then the best bet would be to draw with the percentage of 39.8%.
* If the HTGD is 0 and the FGS is 0, the away team scores and the best bet would be on the away team.
* Node 2: If the HTGD is 2, 3, or 4, we can bet on winning the team with a percentage of 93.1.
* Node 9 and 10: The HTGD is 1 and the Total\_H\_P <= 67, bet on the home team with the percentage of 73.5. and If the Total\_H\_P>= 67, bet with the home team with the percentage of 91.7.
* In Node 11: HTGD is -1 or -4 and the Total\_A\_P <= 53, bet on the away team with the percentage of 46.1.
* In Node 12:HTGD is -1 or -4 and the Total \_A\_P > 53, bet for the away team with the 70.9%

**Question 8**

Based on the Exhibit 10 Inputs calculated multinomial logistic regression to predict the match outcomes

Results:

| Match Number | Match\_O | Odds Wins | Odds  Draw | Prob Win | Prob  loss | Prob  Draw | Prediction |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | -3.561 | -1.526 | 2.19% | 80.38% | 17.47% | Lose |
| 2 | 2 | 3.667 | 3.278 | 54.67% | 1.59% | 42.89% | Win |
| 3 | 2 | -2.573 | -0.949 | -6.52% | 67.39% | 26.09% | Lose |
| 4 | 2 | 3.993 | 3.685 | 57.03% | 1.05% | 41.91% | Win |
| 5 | 1 | 1.076 | 0.189 | 58% | 20% | 24% | Lose |
| 6 | 0 | 1.191 | -0.007 | 62.5% | 19% | 19% | Lose |
| 7 | 0 | 2.253 | 2.601 | 42% | 4.33% | 54.44% | Draw |
| 8 | 2 | 2.481 | 0.823 | 78.35% | 6.61% | 15.05% | Win |
| 9 | 1 | 2.298 | 0.575 | 78.19% | 7.85% | 13.96% | Win |
| 10 | 2 | -1.285 | -0.783 | 14.66% | 58.57% | 26.77% | Lose |
| 11 | 0 | -0.429 | -0.644 | 30.38% | 45.64% | 23.97% | Lose |
| 12 | 2 | 4.032 | 1.222 | 92.78% | 1.64% | 5.57% | Win |
| 13 | 0 | 0.025 | -0.12 | 35.21% | 34.34% | 30.45% | Win |
| 14 | 0 | -3.483 | -1.468 | 2.44% | 79.30% | 18.27% | Lose |
| 15 | 1 | 2.090 | 2.765 | 32.45% | 4.00% | 63.55% | Draw |
| 16 | 2 | 3.719 | 1.238 | 90.23% | 2.20% | 7.57% | Win |
| 17 | 2 | 3.123 | 3.129 | 48.65% | 2.15% | 49.19% | Draw |
| 18 | 0 | -3.245 | -1.554 | 3.11% | 79.98% | 16.91% | Lose |
| 19 | 1 | 2.745 | 2.981 | 43.72% | 2.72% | 53.56% | Draw |
| 20 | 0 | -3.085 | -1.104 | 3.33% | 72.60% | 24.07% | Lose |

**Question 9**

* Chaid takes the outcomes which have higher probability.

| Match No. | Match\_O | Prob Winning | Prob Drawing | Prob Losing | Prediction | CHAID Node | Prediction | Comparison |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | 2.17% | 17.47% | 80.36% | Lose | 5 | Lose | Yes |
| 2 | 2 | 54.91% | 43.41% | 1.68% | Win | 6 | Draw | No |
| 3 | 2 | 6.52% | 26.09% | 67.39% | Lose | 11 | Lose | Yes |
| 4 | 2 | 57.03% | 41.91% | 1.05% | Win | 6 | Draw | No |
| 5 | 1 | 57.05% | 23.5% | 19.45% | Lose | 11 | Lose | Yes |
| 6 | 0 | 62.28% | 18.8% | 18.93% | Lose | 11 | Lose | Yes |
| 7 | 0 | 41.23% | 54.44% | 4.33% | Draw | 6 | Draw | Yes |
| 8 | 2 | 78.35% | 15.05% | 6.61% | Win | 9 | Win | Yes |
| 9 | 1 | 78.19% | 13.96% | 7.85% | Win | 10 | Win | Yes |
| 10 | 2 | 14.66% | 26.77% | 58.57% | Lose | 11 | Lose | Yes |
| 11 | 0 | 30.38% | 23.97% | 45.64% | Lose | 8 | Lose | No |
| 12 | 2 | 92.78% | 5.57% | 1.64% | Win | 2 | Win | Yes |
| 13 | 0 | 35.21% | 30.45% | 34.34% | Win | 7 | Draw | Yes |
| 14 | 0 | 2.44% | 18.27% | 79.30% | Lose | 5 | Lose | Yes |
| 15 | 1 | 32.45% | 63.55% | 4% | Draw | 6 | Draw | Yes |
| 16 | 2 | 90.23% | 7.57% | 2.2% | Win | 2 | Win | Yes |
| 17 | 2 | 48.65% | 49.19% | 2.15% | Draw | 6 | Draw | Yes |
| 18 | 0 | 3.11% | 16.91% | 79.98% | Lose | 5 | Lose | Yes |
| 19 | 1 | 43.72% | 53.56% | 2.72% | Draw | 6 | Draw | Yes |
| 20 | 0 | 3.33% | 24.07% | 72.60% | Lose | 5 | Lose | Yes |

* The model has predicted right for 17/20 matches.

**Question 10**

Peter should bet on the home team in the 12th match because the multinomial outcome predicted they would win and the match resulted in the home team winning.

The CHAIDS decision tree supports this as well, following the HTGD tree and splitting on the difference being 2 indicating that the home team is more likely to win.

Match 12 is the one Peter should bet on since it has the biggest percentage of the home team winning.