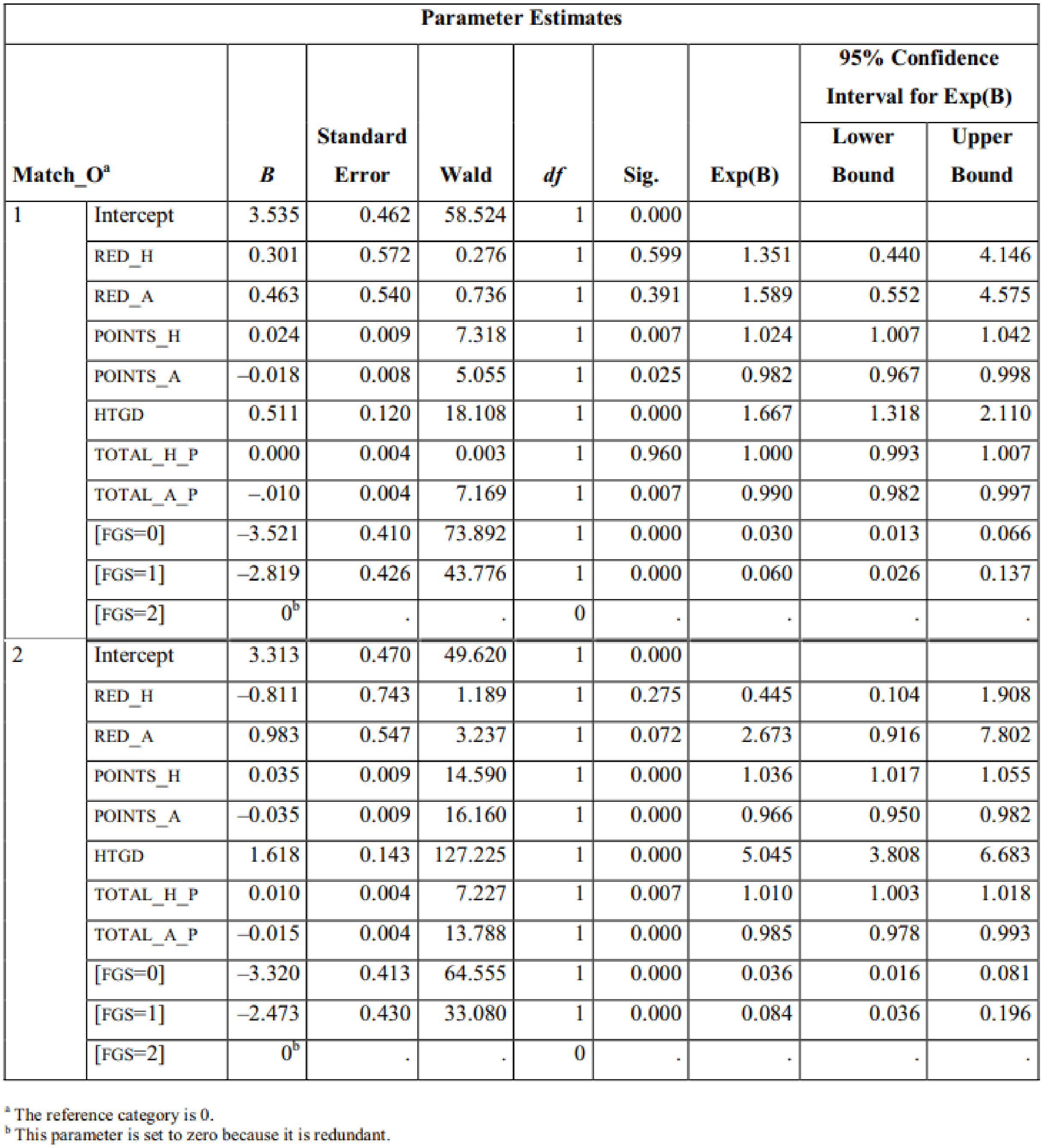
IDS 572 – Data Mining

Home Work 4: A Game of two Halves: In-Play betting in Football

Team Members

Name	UIN
Karthick Sharan	654898680
PraveenKumar V.R.	661639708
Raahul V.S.	678481689

1. Write the equation Peter can use for predicting win for home team using α=0.05.



Parameter Estimates

Match_O[a]		B	Standard Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1	Intercept	3.535	0.462	58.524	1	0.000			
	RED_H	0.301	0.572	0.276	1	0.599	1.351	0.440	4.146
	RED_A	0.463	0.540	0.736	1	0.391	1.589	0.552	4.575
	POINTS_H	0.024	0.009	7.318	1	0.007	1.024	1.007	1.042
	POINTS_A	–0.018	0.008	5.055	1	0.025	0.982	0.967	0.998
	HTGD	0.511	0.120	18.108	1	0.000	1.667	1.318	2.110
	TOTAL_H_P	0.000	0.004	0.003	1	0.960	1.000	0.993	1.007
	TOTAL_A_P	–.010	0.004	7.169	1	0.007	0.990	0.982	0.997
	[FGS=0]	–3.521	0.410	73.892	1	0.000	0.030	0.013	0.066
	[FGS=1]	–2.819	0.426	43.776	1	0.000	0.060	0.026	0.137
	[FGS=2]	0[b]	.	.	0	.	.	.	.
2	Intercept	3.313	0.470	49.620	1	0.000			
	RED_H	–0.811	0.743	1.189	1	0.275	0.445	0.104	1.908
	RED_A	0.983	0.547	3.237	1	0.072	2.673	0.916	7.802
	POINTS_H	0.035	0.009	14.590	1	0.000	1.036	1.017	1.055
	POINTS_A	–0.035	0.009	16.160	1	0.000	0.966	0.950	0.982
	HTGD	1.618	0.143	127.225	1	0.000	5.045	3.808	6.683
	TOTAL_H_P	0.010	0.004	7.227	1	0.007	1.010	1.003	1.018
	TOTAL_A_P	–0.015	0.004	13.788	1	0.000	0.985	0.978	0.993
	[FGS=0]	–3.320	0.413	64.555	1	0.000	0.036	0.016	0.081
	[FGS=1]	–2.473	0.430	33.080	1	0.000	0.084	0.036	0.196
	[FGS=2]	0[b]	.	.	0	.	.	.	.

[a] The reference category is 0.

[b] This parameter is set to zero because it is redundant.

From the given data we know that,

2: Home team won
1: Draw
0: Away team won

And the reference class for the logistic regressor in 0 (Away team won).

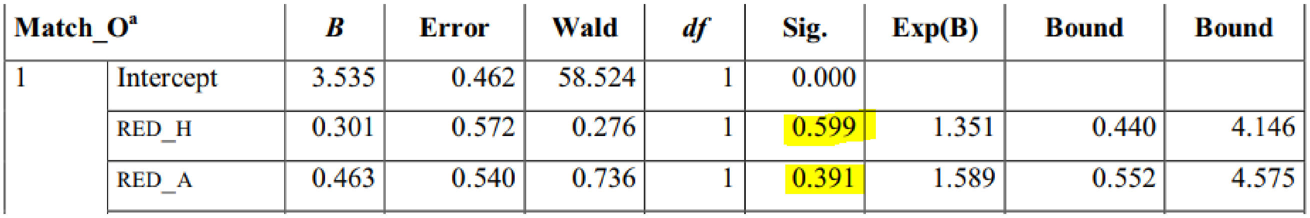
Therefore,
Natural Log of Probability of Home win with respect to away team winning is the equation is:
Ln(P(2)/P(0)) = $Beta_20 + Beta_21*x1 + Beta_22*x2 + ... + Beta_2n*xn$ (for all significant variables)
Ln(P(2)/P(0)) = 3.313 + 0.035*POINT_H + (-0.035)*POINT_A + 1.618*HTGD + 0.010*TOTAL_H_P + (-0.015)*TOTAL_A_P + (-3.320)*FGS_0 + (-2.473)*FGS_1

Similarly for Draw,
Ln(P(1)/P(0)) = $Beta_10 + Beat_11*x1 + Beta_12*x2 + ... + Beta_1n*xn$
= 3.535 + 0.024*POINT_H + (-0.018)*POINT_A + 0.511*HTGD + (-0.010)*TOTAL_A_P + (-3.521)*FGS_0 + (-2.819)*FGS_1

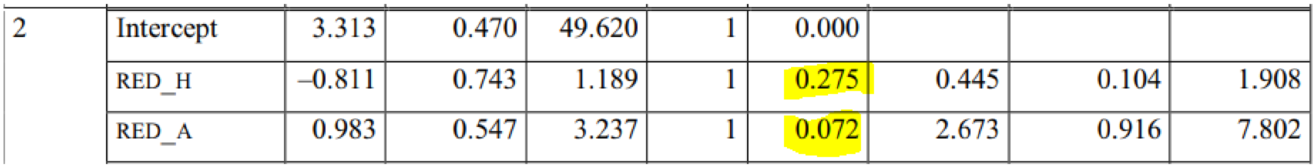
So from the above equations we get, the probability of home team wining as:

$$\boldsymbol{P}(\mathbf{2}) = \frac{\mathbf{e}^{-3.313\,+\,0.035*POINT\_H\,+\,-0.035*POINT\_A\,+\,1.618*HTGD\,+\,0.010*TOTAL\_H\_P\,-0.015*TOTAL\_A\_P\,-3.320*FGS\_0\,-2.473*FGS\_1}}{(\mathbf{1}+\ \mathbf{e}^{-3.313\,+\,0.035*POINT_H+\,-0.035*POINT_A+\,1.618*HTGD\,+\,0.010*TOTAL_{H_P}-0.015*TOTAL_{A_P}-3.320*FGS_0-2.473*FGS_1} + e^{3.535\,+\,0.024*POINT\_H\,-0.018*POINT\_A\,+\,0.511*HTGD\,-0.010*TOTAL\_A\_P\,\,-3.521*FGS\_0\,-2.819*FGS\_1})}$$

2. **What is the influence on the match outcome of red cards conceded by the home and away team? Discuss the possible reasons for the empirical evidence from the model.**

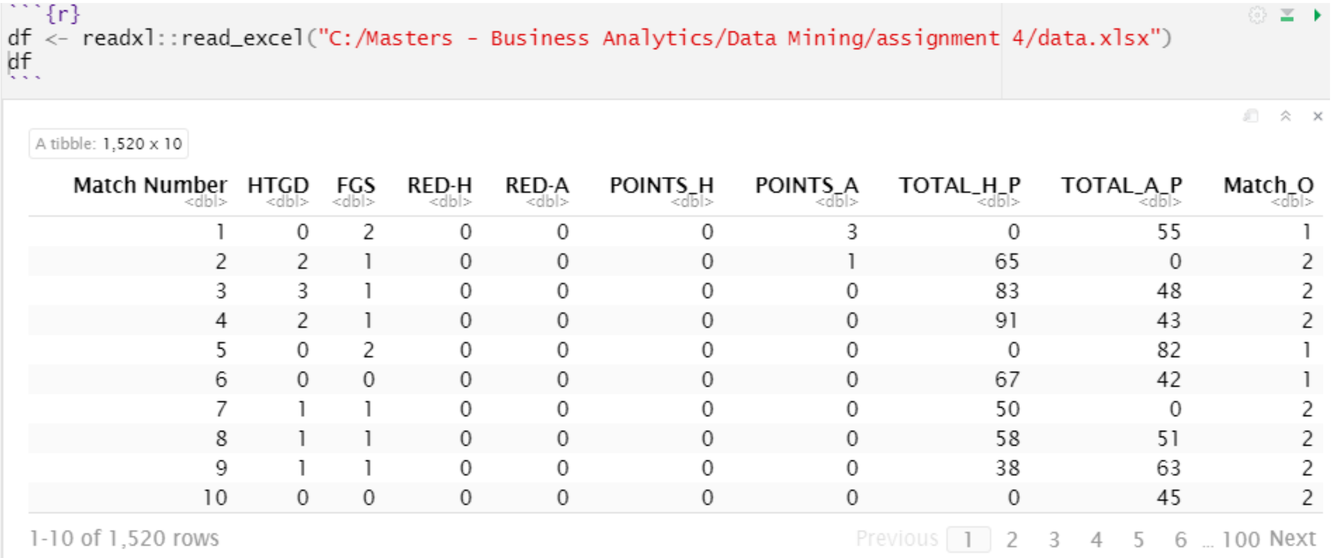


Match_O^a		*B*	Error	Wald	*df*	Sig.	Exp(B)	Bound	Bound
1	Intercept	3.535	0.462	58.524	1	0.000			
	RED_H	0.301	0.572	0.276	1	0.599	1.351	0.440	4.146
	RED_A	0.463	0.540	0.736	1	0.391	1.589	0.552	4.575



2	Intercept	3.313	0.470	49.620	1	0.000			
	RED_H	–0.811	0.743	1.189	1	0.275	0.445	0.104	1.908
	RED_A	0.983	0.547	3.237	1	0.072	2.673	0.916	7.802

For both match outcomes (home win or draw) we can observe that the p-value for Red cards booked to both home and away team is more than the significance limit of 0.05. Hence, **the amount of booked Red cards to either home or away team doesn't impact the match outcome by a considerable amount**.
This means that odds of Home team wining of drawing with respect to away team wining doesn't change whether the home or away team gets booked a red card.



```{r}
df <- readxl::read_excel("C:/Masters - Business Analytics/Data Mining/assignment 4/data.xlsx")
df
```

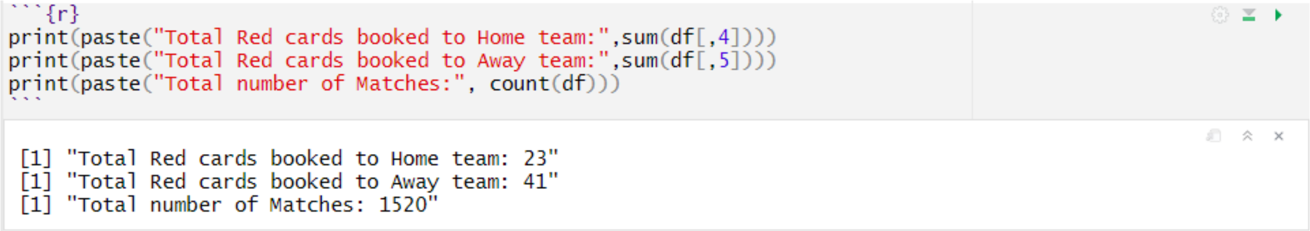
A tibble: 1,520 x 10

Match Number <dbl>	HTGD <dbl>	FGS <dbl>	RED-H <dbl>	RED-A <dbl>	POINTS_H <dbl>	POINTS_A <dbl>	TOTAL_H_P <dbl>	TOTAL_A_P <dbl>	Match_O <dbl>
1	0	2	0	0	0	3	0	55	1
2	2	1	0	0	0	1	65	0	2
3	3	1	0	0	0	0	83	48	2
4	2	1	0	0	0	0	91	43	2
5	0	2	0	0	0	0	0	82	1
6	0	0	0	0	0	0	67	42	1
7	1	1	0	0	0	0	50	0	2
8	1	1	0	0	0	0	58	51	2
9	1	1	0	0	0	0	38	63	2
10	0	0	0	0	0	0	0	45	2

1-10 of 1,520 rows Previous 1 2 3 4 5 6 ... 100 Next

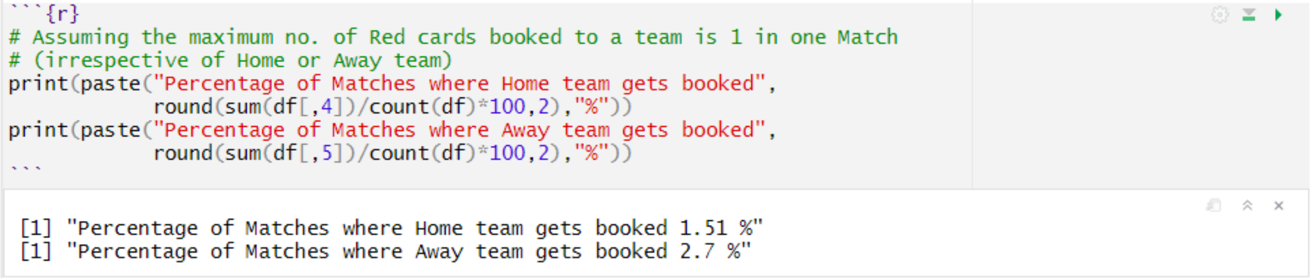
(Loading the dataset of the model was built on in R, to analyze)

The reason why Red card booking aren't affecting the Match outcome according to the model is probably because the number of matches where Red cards were actually given out.



```{r}
print(paste("Total Red cards booked to Home team:",sum(df[,4])))
print(paste("Total Red cards booked to Away team:",sum(df[,5])))
print(paste("Total number of Matches:", count(df)))
```

```
[1] "Total Red cards booked to Home team: 23"
[1] "Total Red cards booked to Away team: 41"
[1] "Total number of Matches: 1520"
```

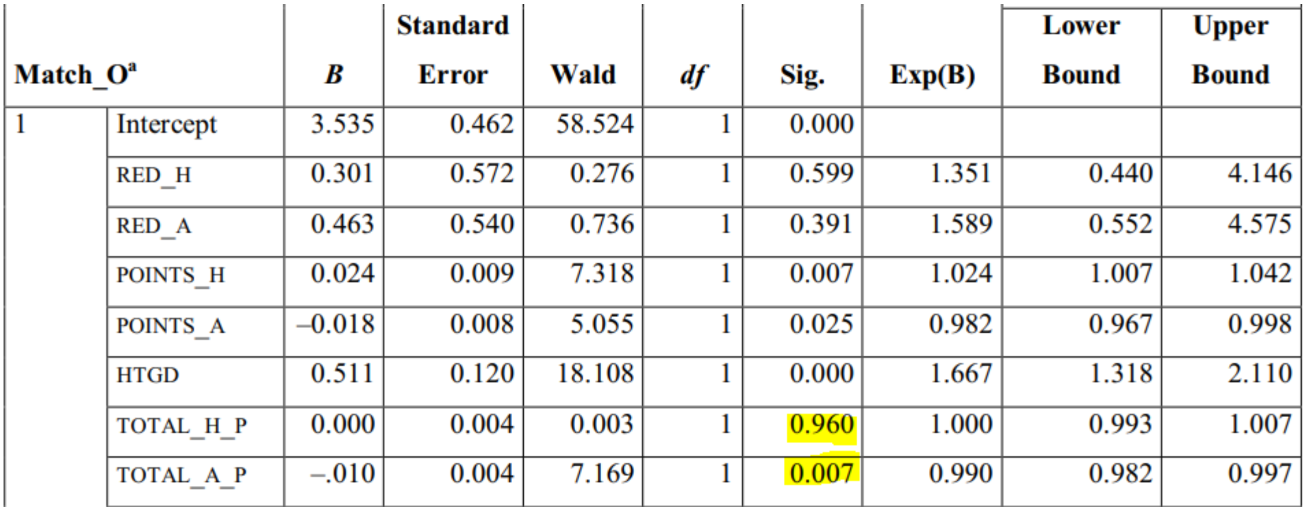


```{r}
# Assuming the maximum no. of Red cards booked to a team is 1 in one Match
# (irrespective of Home or Away team)
print(paste("Percentage of Matches where Home team gets booked",
            round(sum(df[,4])/count(df)*100,2),"%"))
print(paste("Percentage of Matches where Away team gets booked",
            round(sum(df[,5])/count(df)*100,2),"%"))
```

```
[1] "Percentage of Matches where Home team gets booked 1.51 %"
[1] "Percentage of Matches where Away team gets booked 2.7 %"
```

From the above output we can see that out the total records only 1.5% matches saw Red cards handed out to home team and in only 2.7% of the matches the Away team got a Red card. So almost in ~98% of the total matches played no team got a red card, so the number of matches are too low to draw out a relation with respect to the target. So, we can say that in both the RED_H and RED_A column the value throughout the column is same(0) irrespective of the observation. Hence**, these 2 are zero variance columns and they don't add any useful information to the model.**

3. Is it relevant to use the points scored by a team in the previous season for predicting the outcome of a match?

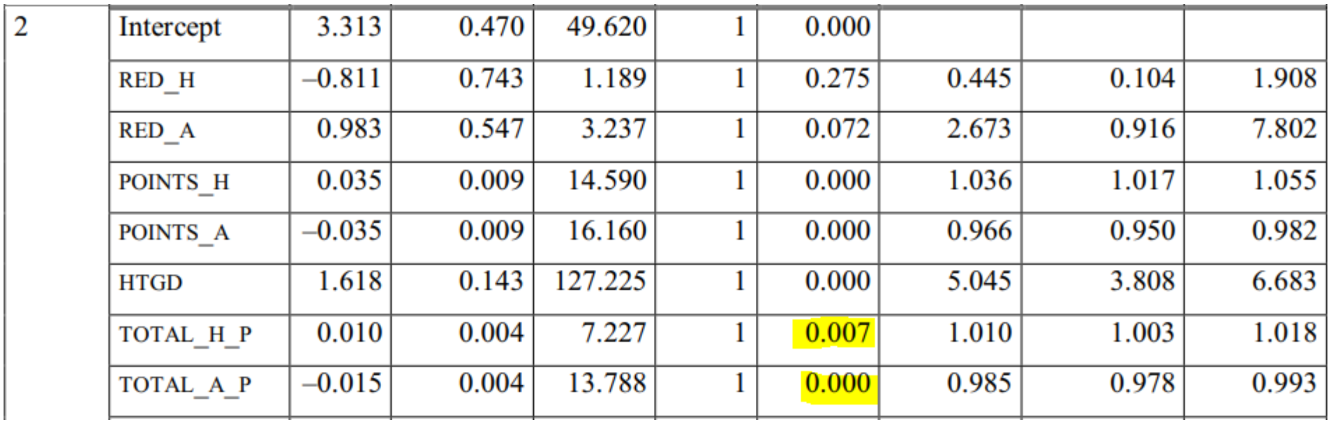


Match_O[a]		*B*	Standard Error	Wald	*df*	Sig.	Exp(B)	Lower Bound	Upper Bound
1	Intercept	3.535	0.462	58.524	1	0.000			
	RED_H	0.301	0.572	0.276	1	0.599	1.351	0.440	4.146
	RED_A	0.463	0.540	0.736	1	0.391	1.589	0.552	4.575
	POINTS_H	0.024	0.009	7.318	1	0.007	1.024	1.007	1.042
	POINTS_A	–0.018	0.008	5.055	1	0.025	0.982	0.967	0.998
	HTGD	0.511	0.120	18.108	1	0.000	1.667	1.318	2.110
	TOTAL_H_P	0.000	0.004	0.003	1	0.960	1.000	0.993	1.007
	TOTAL_A_P	–.010	0.004	7.169	1	0.007	0.990	0.982	0.997

From the Model summary we can see that for outcome1 (Draw), with respect to outcome0 (Away win), TOTAL_H_P (total points scored by home team in previous season) is not significant and doesn't impact the outcome.
However, we can also observe that the total points score by away team in the previous league (TOTAL_A_P) is in fact significant and has a negative co-efficient.
So, we infer that if all other parameters are set than for every 1 point increase in the total points of away team scored in previous season, the odds of a draw decreases by a factor of (1-$e^{-0.01}$)

In other words, for every 1 point increase in away team's previous season total, the odds of a Draw decreases by 1-$e^{-0.01}$ = 0.01 ~ 1%.

For outcome2 (Home team winning)

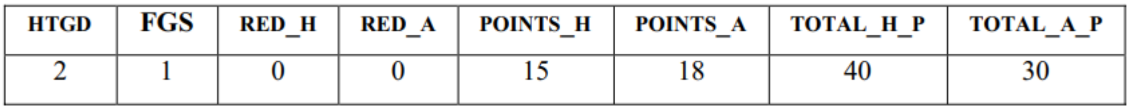


2	Intercept	3.313	0.470	49.620	1	0.000			
	RED_H	–0.811	0.743	1.189	1	0.275	0.445	0.104	1.908
	RED_A	0.983	0.547	3.237	1	0.072	2.673	0.916	7.802
	POINTS_H	0.035	0.009	14.590	1	0.000	1.036	1.017	1.055
	POINTS_A	–0.035	0.009	16.160	1	0.000	0.966	0.950	0.982
	HTGD	1.618	0.143	127.225	1	0.000	5.045	3.808	6.683
	TOTAL_H_P	0.010	0.004	7.227	1	0.007	1.010	1.003	1.018
	TOTAL_A_P	–0.015	0.004	13.788	1	0.000	0.985	0.978	0.993

Here we can observe that, total points scored in previous season affects the match outcome.
Given all other parameters are set, for every 1 point increase in previous season's points of home team, the odd of home team winning increases by a factor of $e^{0.01}$, which is approximately a 1% increase in odds.
Similarly, for 1 point increase in away team's previous season score the odds of home team wining decreases by a factor of 1-$e^{-.015}$=0.015, or we can say the odds of home team wining decreased by approximately 1.5%

From the above 2 scenario we can infer **that the total points scored by a team in the previous season does impact the match outcome, but given the competitive nature of EPL most team finish around the same range with little difference in points. So, unless the point difference between the two teams are considerably high, there are better variable to base our predictions on.**

4. **What is the probability that the home team will win the match for the values below:**



HTGD	FGS	RED_H	RED_A	POINTS_H	POINTS_A	TOTAL_H_P	TOTAL_A_P
2	1	0	0	15	18	40	30

From the equation in question 1, substituting the values we get the (considering only significant variables):

$$P(2) = \frac{e^{3.313+1.618*2-2.473*1+0.035*15-0.035*18+0.01*40-0.015*30}}{1+e^{3.313+1.618*2-2.473*1+0.035*15-0.035*18+0.01*40-0.015*30}+e^{3.535+0.511*2-2.819*1+0.024*15-0.018*18-0.01*30}}$$

$$= \frac{50.45}{(1+50.45+4.367)}$$

$$= 0.90385$$

Therefore, the **probability of home team winning is 0.90385 or 90.385%.**

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 other variables in the regression model.**

From the Model summary we can observe that the co-efficient for FGS_0 has the highest negative value for both outcomes Home-win and Draw, hence it has a considerable amount of impact on reducing the odds for home win/draw.
From the model summary, if the away team scored the first goal scored, then the odds of Draw decreases by a factor of (1-$e^{-3.521}$) with respect to no goals scored, provided all other parameters are set. Similarly, the odds of home team decrease by a factor of (1-$e^{-3.320}$) % w.r.t to no goals scored.

For the above data, if FGS = 0 instead of 1, then the new probability of home win is:

$$P(2) = \frac{21.63}{1+21.63+2.164} = 0.8724$$

So the chances of Home team winning decreases by nearly 3% if all other variables are the same and if the first goal was scored by the away team.

Now considering if the away team scores the first goal, assuming no other goals went in the first half, the value for FGS is 0, HTGD is -1, (assuming both teams have equal point for the current and previous season):

$$P(1) = \frac{1}{1+e^{3.535+0.024*15-0.018*15+0.511*-1-0.01*30-3.521*1}+e^{3.313+0.035*15-0.035*15+1.618*-1+0.01*30-0.015*30-3.320*1}}$$

$$= \frac{1}{(1+0.4931+0.1695)} = 0.6015$$

Now if, the away team has performed better than the home team in the current and previous league: (considering the total point for the away team is 25 and 50 for current and previous year respectively, other variables remain the same)

$$P(1) = \frac{1}{1+e^{3.535+0.024*15-0.018*25+0.511*-1-0.01*50-3.521*1}+e^{3.313+0.035*15-0.035*25+1.618*-1+0.01*30-0.015*50-3.320*1}}$$

$$= \frac{1}{(1+0.337+0.088)} = 0.7017$$

Lastly, considering away team scored more than 1 goals in the first half, (HTGD = -3):

$$P(1) = \frac{1}{1+e^{3.535+0.024*15-0.018*25+0.511*-3-0.01*50-3.521*1}+e^{3.313+0.035*15-0.035*25+1.618*-3+0.01*30-0.015*50-3.320*1}}$$

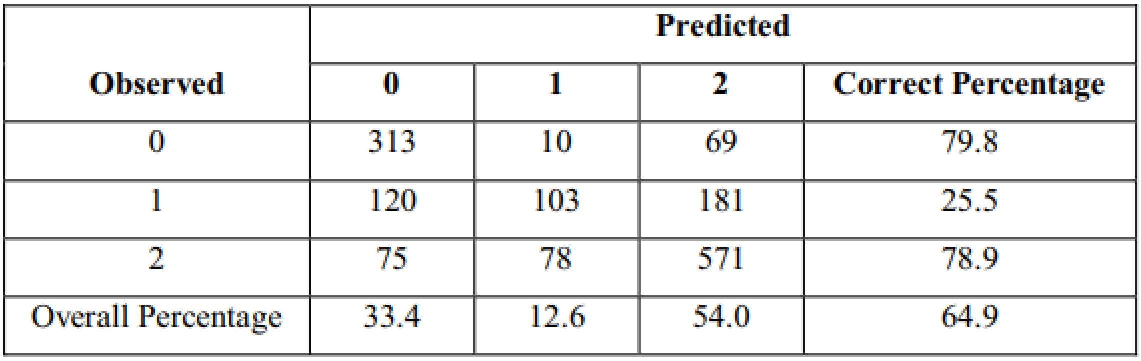
$$= \frac{1}{(1+0.1213+0.0034)} = 0.8891$$

From the above calculation we can observe that if **two equally good teams are playing** and if the away team scores the first goal then **the probability of the away team winning is 0.6015 or nearly 60.15%.** In case the **away team is a better one** then as per our assumptions the probability of **away team winning is 0.7017 or ~70%.**
If the **away team has scored more than 1 (3 goals)** in the first half then **probability increases to nearly 90%.**

So, to conclude if the teams are somewhat on the same tier then while the away team scoring the first does impact the outcome and can be betted on, however it's also a risky option since the probability of them winning is only 0.6015. In combination with the FGS if we check other factors as well like the halftime goal difference and difference in the current league standing of the 2 teams then we can get a much more comprehensive picture and can make a much more safer bet.

6. **What conclusion can you derive from the classification table, is it advisable to bet on Draws?**

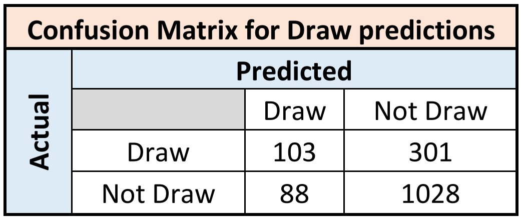
We have the below confusion matrix from the built model:



Observed	Predicted			
	0	1	2	Correct Percentage
0	313	10	69	79.8
1	120	103	181	25.5
2	75	78	571	78.9
Overall Percentage	33.4	12.6	54.0	64.9

Based on the confusion matrix, we can draw the below conclusions:
- The Recall for Away team wining is 79.8%, and for home team wining its 78.9%. Hence our model performs decently well for these two outcomes.
- Precision for Away team winning is 61.61% and the precision for Home team wins are nearly 70%.
- From the Recall and Precision values we can infer that most of the Matches which actually ended in a Draw are being categorized as either Away win or Home win.

To further analyze the model with respect to Draw matches we can reduce the above confusion matrix to a binary table, shown below:



Confusion Matrix for Draw predictions

		Predicted	
Actual		Draw	Not Draw
	Draw	103	301
	Not Draw	88	1028

From the confusion table above we can calculate the Recall, precision for Draw matches (outcome 1 in the original matrix):
- As expected the Recall for Draw matches predicted by the model is very low at only around **25.5%** (103/404).
- The precision for draw matches aren't that good either with only **54%** (103/191).
  Hence the obtained **f-score** is **34.6%** (2*Recall*Precision)/(Recall+Precision).

So based on the above observations we can conclude that the model performs worse for matches which end in a Draw, with very low f-score, hence **it's strongly advised not to bet on Draws** using this Model

7. Using the CHAID decision tree shown in Exhibit 9, frame rules that may be used for betting.

Using the CHAID decision tree, we can observe below 3 categories of decision rules:

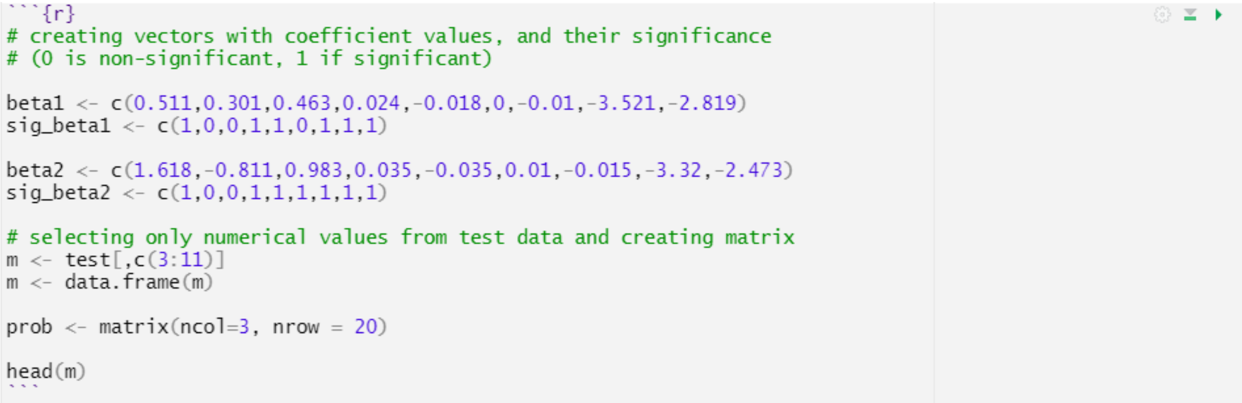
<u>Strong Decision tree rules with confidence over 90%:</u>
HTGD = 2.0, 3.0, 4.0 → 2 (Home team wins) [Confidence = 93.1%, Support = 11.4%]
HTGD = -2.0, -3.0, -5.0 → 0 (Away team wins) [Confidence = 91.3%, Support = 4.5%]
HTGD = 1.0; Total_H_P > 67.0 → 2 (Home team wins) [Confidence = 91.7%, Support = 5.5%]

<u>Decision tree rules with an Confidence >70% and <90%:</u>
HTGD = 1.0; Total_H_P <= 67.0 → 2 (Home team wins) [Confidence = 73.5%, Support = 19.3%]
HTGD = -1.0, -4.0; Total_A_P > 53.0 → 0 (Away team wins) [Confidence = 70.9%, Support = 9.3%]

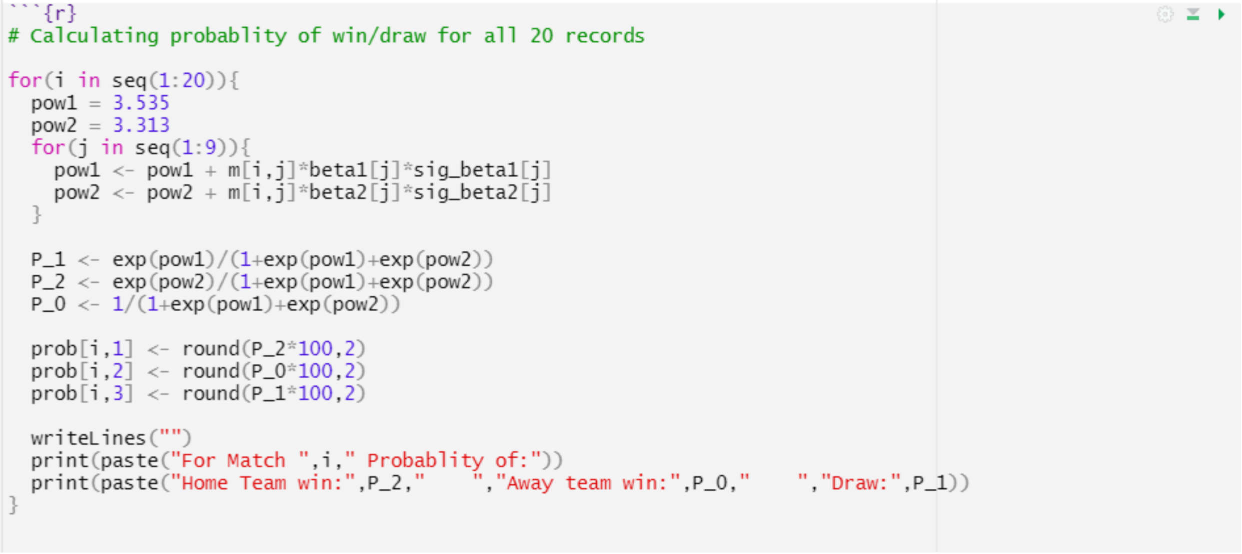
<u>Weak Decision tree rules with bad Confidence < 50%:</u>
HTGD = 0.0; FGS = 2.0 → 1 (Draw) [Confidence = 49.7%, Support = 20.7%]
HTGD = 0.0; FGS = 1.0 → 1 (Draw) [Confidence = 39.8%, Support = 6.1%]
HTGD = 0.0; FGS = 0.0 → 0 (Away team wins) [Confidence = 49.1%, Support = 14.7%]
HTGD = -1.0, -4.0; Total_A_P <= 53.0 → 0 (Away team wins) [Confidence = 46.1%, Support = 8.4%]

8. Use multinomial Regression to predict the match outcome in all 20 cases listed.

The following code is used to calculate the probability for the match outcomes for the 20 test records:



~~~~
```{r}
# creating vectors with coefficient values, and their significance
# (0 is non-significant, 1 if significant)

beta1 <- c(0.511,0.301,0.463,0.024,-0.018,0,-0.01,-3.521,-2.819)
sig_beta1 <- c(1,0,0,1,1,0,1,1,1)

beta2 <- c(1.618,-0.811,0.983,0.035,-0.035,0.01,-0.015,-3.32,-2.473)
sig_beta2 <- c(1,0,0,1,1,1,1,1,1)

# selecting only numerical values from test data and creating matrix
m <- test[,c(3:11)]
m <- data.frame(m)

prob <- matrix(ncol=3, nrow = 20)

head(m)
```
~~~~



~~~~
```{r}
# Calculating probablity of win/draw for all 20 records

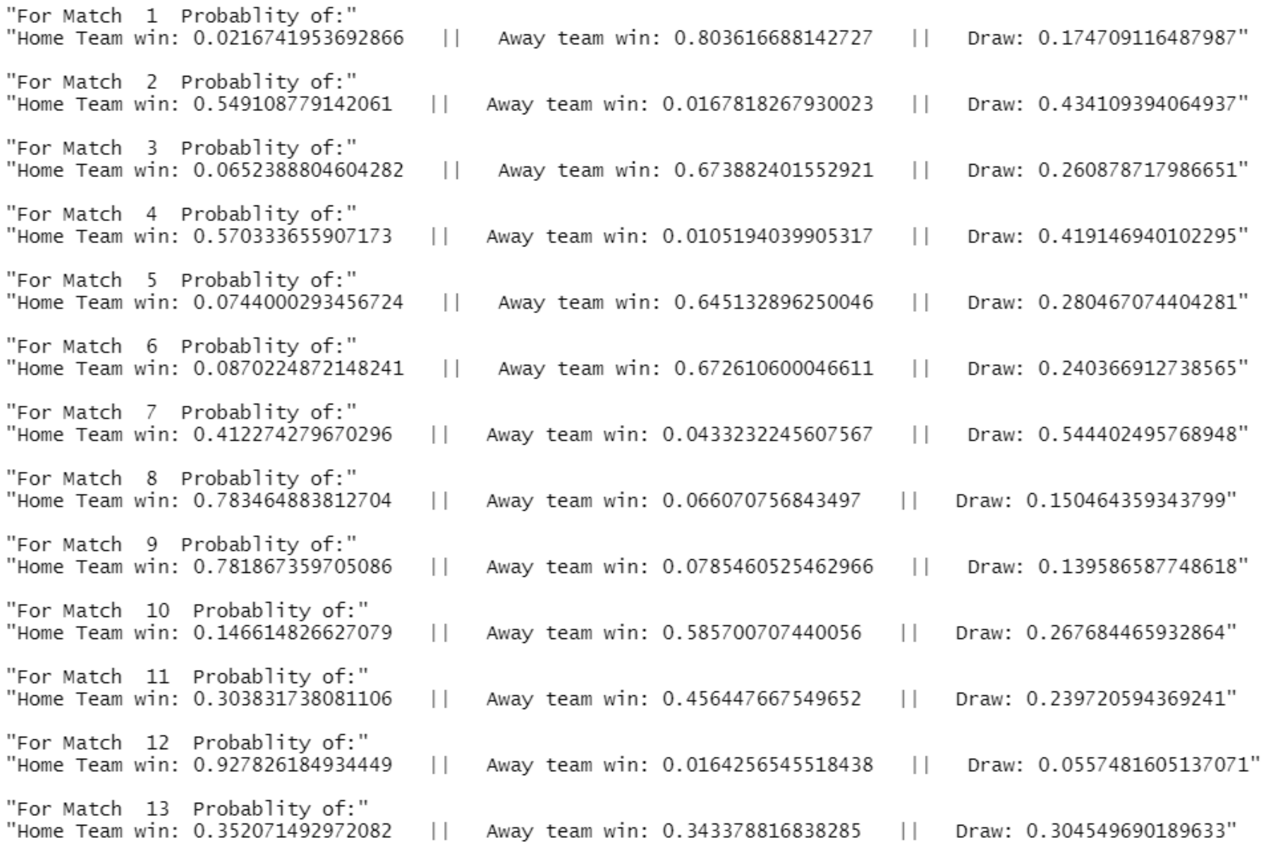
for(i in seq(1:20)){
  pow1 = 3.535
  pow2 = 3.313
  for(j in seq(1:9)){
    pow1 <- pow1 + m[i,j]*beta1[j]*sig_beta1[j]
    pow2 <- pow2 + m[i,j]*beta2[j]*sig_beta2[j]
  }

  P_1 <- exp(pow1)/(1+exp(pow1)+exp(pow2))
  P_2 <- exp(pow2)/(1+exp(pow1)+exp(pow2))
  P_0 <- 1/(1+exp(pow1)+exp(pow2))

  prob[i,1] <- round(P_2*100,2)
  prob[i,2] <- round(P_0*100,2)
  prob[i,3] <- round(P_1*100,2)

  writeLines("")
  print(paste("For Match ",i," Probablity of:"))
  print(paste("Home Team win:",P_2,"    ","Away team win:",P_0,"    ","Draw:",P_1))
}
~~~~

Output for the above code :



```
"For Match  1  Probablity of:"
"Home Team win: 0.0216741953692866   ||   Away team win: 0.803616688142727   ||   Draw: 0.174709116487987"

"For Match  2  Probablity of:"
"Home Team win: 0.549108779142061   ||   Away team win: 0.0167818267930023   ||   Draw: 0.434109394064937"

"For Match  3  Probablity of:"
"Home Team win: 0.0652388804604282   ||   Away team win: 0.673882401552921   ||   Draw: 0.260878717986651"

"For Match  4  Probablity of:"
"Home Team win: 0.570333655907173   ||   Away team win: 0.0105194039905317   ||   Draw: 0.419146940102295"

"For Match  5  Probablity of:"
"Home Team win: 0.0744000293456724   ||   Away team win: 0.645132896250046   ||   Draw: 0.280467074404281"

"For Match  6  Probablity of:"
"Home Team win: 0.0870224872148241   ||   Away team win: 0.672610600046611   ||   Draw: 0.240366912738565"

"For Match  7  Probablity of:"
"Home Team win: 0.412274279670296   ||   Away team win: 0.0433232245607567   ||   Draw: 0.544402495768948"

"For Match  8  Probablity of:"
"Home Team win: 0.783464883812704   ||   Away team win: 0.066070756843497   ||   Draw: 0.150464359343799"

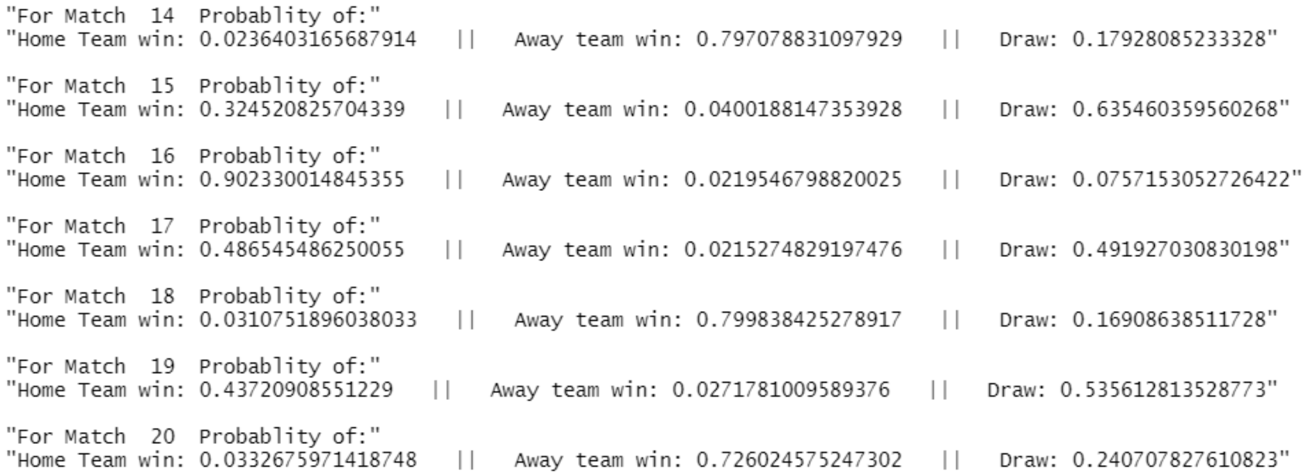
"For Match  9  Probablity of:"
"Home Team win: 0.781867359705086   ||   Away team win: 0.0785460525462966   ||   Draw: 0.139586587748618"

"For Match  10  Probablity of:"
"Home Team win: 0.146614826627079   ||   Away team win: 0.585700707440056   ||   Draw: 0.267684465932864"

"For Match  11  Probablity of:"
"Home Team win: 0.303831738081106   ||   Away team win: 0.456447667549652   ||   Draw: 0.239720594369241"

"For Match  12  Probablity of:"
"Home Team win: 0.927826184934449   ||   Away team win: 0.0164256545518438   ||   Draw: 0.0557481605137071"

"For Match  13  Probablity of:"
"Home Team win: 0.352071492972082   ||   Away team win: 0.343378816838285   ||   Draw: 0.304549690189633"
```



```
"For Match  14  Probablity of:"
"Home Team win: 0.0236403165687914   ||   Away team win: 0.797078831097929   ||   Draw: 0.17928085233328"

"For Match  15  Probablity of:"
"Home Team win: 0.324520825704339   ||   Away team win: 0.0400188147353928   ||   Draw: 0.635460359560268"

"For Match  16  Probablity of:"
"Home Team win: 0.902330014845355   ||   Away team win: 0.0219546798820025   ||   Draw: 0.0757153052726422"

"For Match  17  Probablity of:"
"Home Team win: 0.486545486250055   ||   Away team win: 0.0215274829197476   ||   Draw: 0.491927030830198"

"For Match  18  Probablity of:"
"Home Team win: 0.0310751896038033   ||   Away team win: 0.799838425278917   ||   Draw: 0.16908638511728"

"For Match  19  Probablity of:"
"Home Team win: 0.43720908551229   ||   Away team win: 0.0271781009589376   ||   Draw: 0.535612813528773"

"For Match  20  Probablity of:"
"Home Team win: 0.0332675971418748   ||   Away team win: 0.726024575247302   ||   Draw: 0.240707827610823"
```

So, after converting the probability to classes we can get the following prediction for the 20 matches:
(To convert the outcome which has the highest probability among the 3 is chosen as final outcome)

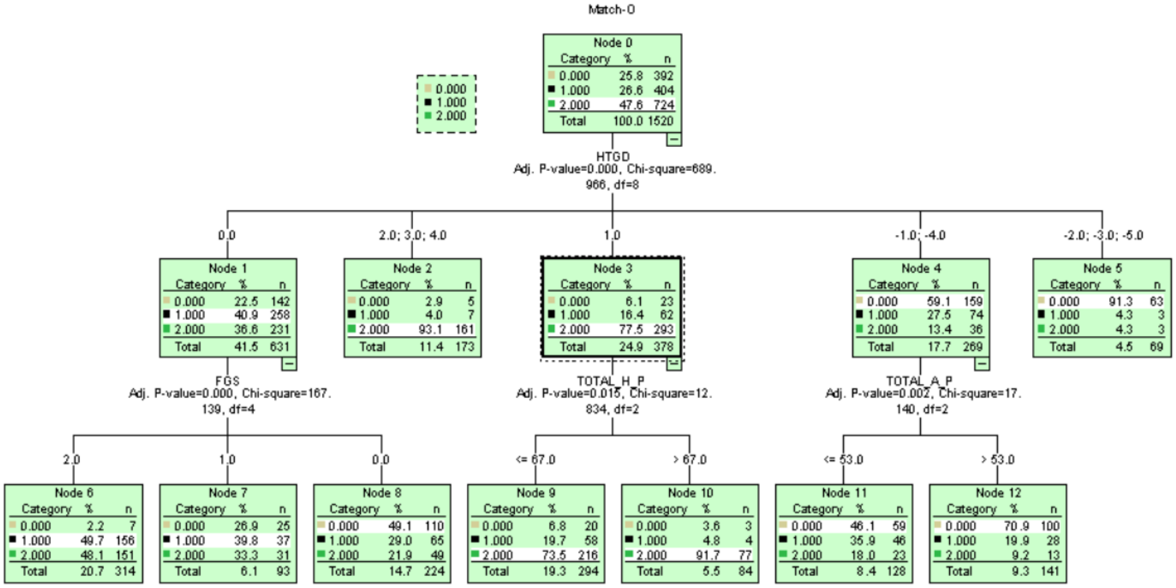
Match Number	Match betweenw	MATCH_O	Predicted (Logistic Regressor)
1	SwanseaCity vs. Everton	0	**0**
2	Chelsea vs. StokeCity	2	**2**
3	Southampton vs. AstonVilla	2	**0**
4	WestBromwichAlbion vs. Reading	2	**2**
5	WestHamUnited vs. Sunderland	1	**0**
6	WiganAthletic vs. Fulham	0	**0**
7	Liverpool vs. ManchesterUnited	0	**1**
8	NewcastleUnited vs. NorwichCity	2	**2**
9	ManchesterCity vs. Arsenal	1	**2**
10	TottenhamHotspur vs. QueensParkRangers	2	**0**
11	Arsenal vs. Chelsea	0	**0**
12	Everton vs. Southampton	2	**2**
13	Fulham vs. ManchesterCity	0	**2**
14	NorwichCity vs. Liverpool	0	**0**
15	Reading vs. NewcastleUnited	1	**1**
16	StokeCity vs. SwanseaCity	2	**2**
17	Sunderland vs. WiganAthletic	2	**1**
18	ManchesterUnited vs. TottenhamHotspurs	0	**0**
19	AstonVilla vs. WestBromwichAlbion	1	**1**
20	QueensParkRangers vs. WestHam	0	**0**

The Matches highlighted in orange are the one which are misclassified by the provided logistic regression model.

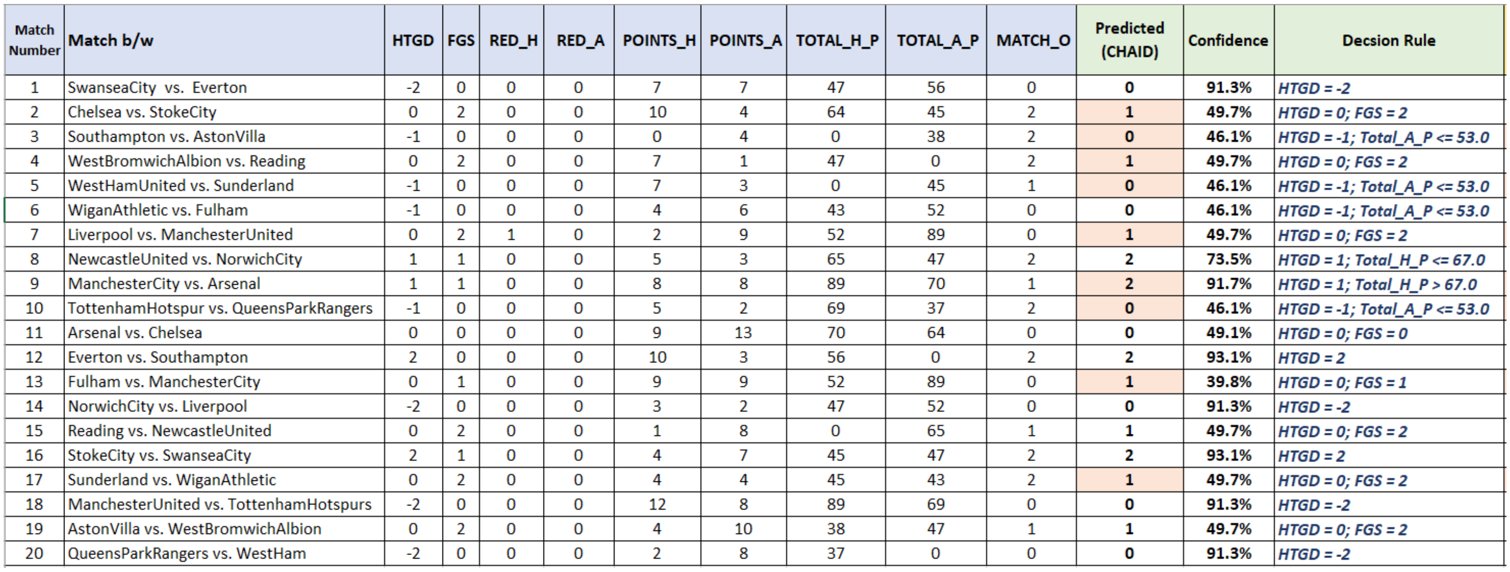
We can see that there a quite a few misclassification, and the **accuracy of the Regressor is 65%.**

9. **Apply the CHAID decision tree on the 20 matches and compare the results with your answers obtained using multinomial logistic regression.**

The provided decision tree is:
Using the above decision tree we can predict the outcomes for the below 20 test cases(Matches):



Prediction of Match outcome using decision tree:



Match Number	Match b/w	HTGD	FGS	RED_H	RED_A	POINTS_H	POINTS_A	TOTAL_H_P	TOTAL_A_P	MATCH_O	Predicted (CHAID)	Confidence	Decsion Rule
1	SwanseaCity vs. Everton	-2	0	0	0	7	7	47	56	0	**0**	**91.3%**	***HTGD = -2***
2	Chelsea vs. StokeCity	0	2	0	0	10	4	64	45	2	**1**	**49.7%**	***HTGD = 0; FGS = 2***
3	Southampton vs. AstonVilla	-1	0	0	0	0	4	0	38	2	**0**	**46.1%**	***HTGD = -1; Total_A_P <= 53.0***
4	WestBromwichAlbion vs. Reading	0	2	0	0	7	1	47	0	2	**1**	**49.7%**	***HTGD = 0; FGS = 2***
5	WestHamUnited vs. Sunderland	-1	0	0	0	7	3	0	45	1	**0**	**46.1%**	***HTGD = -1; Total_A_P <= 53.0***
6	WiganAthletic vs. Fulham	-1	0	0	0	4	6	43	52	0	**0**	**46.1%**	***HTGD = -1; Total_A_P <= 53.0***
7	Liverpool vs. ManchesterUnited	0	2	1	0	2	9	52	89	0	**1**	**49.7%**	***HTGD = 0; FGS = 2***
8	NewcastleUnited vs. NorwichCity	1	1	0	0	5	3	65	47	2	**2**	**73.5%**	***HTGD = 1; Total_H_P <= 67.0***
9	ManchesterCity vs. Arsenal	1	1	0	0	8	8	89	70	1	**2**	**91.7%**	***HTGD = 1; Total_H_P > 67.0***
10	TottenhamHotspur vs. QueensParkRangers	-1	0	0	0	5	2	69	37	2	**0**	**46.1%**	***HTGD = -1; Total_A_P <= 53.0***
11	Arsenal vs. Chelsea	0	0	0	0	9	13	70	64	0	**0**	**49.1%**	***HTGD = 0; FGS = 0***
12	Everton vs. Southampton	2	0	0	0	10	3	56	0	2	**2**	**93.1%**	***HTGD = 2***
13	Fulham vs. ManchesterCity	0	1	0	0	9	9	52	89	0	**1**	**39.8%**	***HTGD = 0; FGS = 1***
14	NorwichCity vs. Liverpool	-2	0	0	0	3	2	47	52	0	**0**	**91.3%**	***HTGD = -2***
15	Reading vs. NewcastleUnited	0	2	0	0	1	8	0	65	1	**1**	**49.7%**	***HTGD = 0; FGS = 2***
16	StokeCity vs. SwanseaCity	2	1	0	0	4	7	45	47	2	**2**	**93.1%**	***HTGD = 2***
17	Sunderland vs. WiganAthletic	0	2	0	0	4	4	45	43	2	**1**	**49.7%**	***HTGD = 0; FGS = 2***
18	ManchesterUnited vs. TottenhamHotspurs	-2	0	0	0	12	8	89	69	0	**0**	**91.3%**	***HTGD = -2***
19	AstonVilla vs. WestBromwichAlbion	0	2	0	0	4	10	38	47	1	**1**	**49.7%**	***HTGD = 0; FGS = 2***
20	QueensParkRangers vs. WestHam	-2	0	0	0	2	8	37	0	0	**0**	**91.3%**	***HTGD = -2***

The above table displays all the 20 records for which we need to predict the outcome.

Using the decision rules from the above tree we can predict the classes of the outcome, in the above table the outcome of matches predicted using the CHAID decision tree is shown along with the confidence for that prediction and the applicable rules for that particular match(observation).

We can notice that there are **9 matches were the predictions were wrong**, hence the overall **accuracy of the model is only 55%.**

By comparing both models we get the following results:

Match Number	Match b/w	MATCH_O	Predicted (Logistic Regressor)	Predicted (CHAID)
1	SwanseaCity vs. Everton	0	**0**	**0**
2	Chelsea vs. StokeCity	2	**2**	**1**
3	Southampton vs. AstonVilla	2	**0**	**0**
4	WestBromwichAlbion vs. Reading	2	**2**	**1**
5	WestHamUnited vs. Sunderland	1	**0**	**0**
6	WiganAthletic vs. Fulham	0	**0**	**0**
7	Liverpool vs. ManchesterUnited	0	**1**	**1**
8	NewcastleUnited vs. NorwichCity	2	**2**	**2**
9	ManchesterCity vs. Arsenal	1	**2**	**2**
10	TottenhamHotspur vs. QueensParkRangers	2	**0**	**0**
11	Arsenal vs. Chelsea	0	**0**	**0**
12	Everton vs. Southampton	2	**2**	**2**
13	Fulham vs. ManchesterCity	0	**2**	**1**
14	NorwichCity vs. Liverpool	0	**0**	**0**
15	Reading vs. NewcastleUnited	1	**1**	**1**
16	StokeCity vs. SwanseaCity	2	**2**	**2**
17	Sunderland vs. WiganAthletic	2	**1**	**1**
18	ManchesterUnited vs. TottenhamHotspurs	0	**0**	**0**
19	AstonVilla vs. WestBromwichAlbion	1	**1**	**1**
20	QueensParkRangers vs. WestHam	0	**0**	**0**

Accuracy of Multi class Logistic Regressor	**65%**
Accuracy of CHAID decision tree Model	**55%**

We can see that the **multi-class logistic regressor perform a bit better than the CHAID model. The regressor correctly predicted the outcome for 13 matches, while the CHAID only predicted for 11 matches out of 20.**

Along with all the misclassification of the regressor there are also some extra records which are misclassified in the CHAID decision tree model. Considering only these 20 matches, we can say that **the multinomial regressor is a better model when compared to the CHAID decision tree built for this dataset.**

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

According to the multiclass logistic regressor, the match Peter should bet on is **Match 12**, since it has the highest probability (irrespective of who wins) value among all the Matches. From the model results we can see that for **match 12 it predicts that the home team will win, with a probability of 92.78%.**

The following code snippet can be used to fetch the value and the respective match:

For the below table (probability prediction of logistic regressor):

```
df_prob
```

```
##    Home Win(%) Away Win(%) Draw(%)
## 1         2.17       80.36   17.47
## 2        54.91        1.68   43.41
## 3         6.52       67.39   26.09
## 4        57.03        1.05   41.91
## 5         7.44       64.51   28.05
## 6         8.70       67.26   24.04
## 7        41.23        4.33   54.44
## 8        78.35        6.61   15.05
## 9        78.19        7.85   13.96
## 10       14.66       58.57   26.77
## 11       30.38       45.64   23.97
## 12       92.78        1.64    5.57
## 13       35.21       34.34   30.45
## 14        2.36       79.71   17.93
## 15       32.45        4.00   63.55
## 16       90.23        2.20    7.57
## 17       48.65        2.15   49.19
## 18        3.11       79.98   16.91
## 19       43.72        2.72   53.56
## 20        3.33       72.60   24.07
```

We get the below results:

```
print(paste("Home Win: Match",which.max(df_prob$`Home Win(%)`)," -",
            df_prob[which.max(df_prob$`Home Win(%)`),1],"%"))
```

```
## [1] "Home Win: Match 12  - 92.78 %"
```

```
print(paste("Away Win: Match",which.max(df_prob$`Away Win(%)`)," -",
            df_prob[which.max(df_prob$`Away Win(%)`),2],"%"))
```

```
## [1] "Away Win: Match 1  - 80.36 %"
```

```
print(paste("Draw: Match",which.max(df_prob$`Draw`)," -",
            df_prob[which.max(df_prob$`Draw`),3],"%"))
```

```
## [1] "Draw: Match 15  - 63.55 %"
```

The Highest probability for Home team to win occurs at Match 12, highest probability for away to win is in Match 1, betting on a Draw outcome isn't advisable since the highest probability for a draw is only 63%.

So from the above table we can say that **betting on the home team in Match 12 is the best option, followed by betting on home team in Match 16**

Now, using the output from CHAID decision tree:

Match Number	Match b/w	Predicted (CHAID)	Confidence	Decsion Rule
1	SwanseaCity vs. Everton	**0**	**91.3%**	HTGD = -2
2	Chelsea vs. StokeCity	**1**	**49.7%**	HTGD = 0; FGS = 2
3	Southampton vs. AstonVilla	**0**	**46.1%**	HTGD = -1; Total_A_P <= 53.0
4	WestBromwichAlbion vs. Reading	**1**	**49.7%**	HTGD = 0; FGS = 2
5	WestHamUnited vs. Sunderland	**0**	**46.1%**	HTGD = -1; Total_A_P <= 53.0
6	WiganAthletic vs. Fulham	**0**	**46.1%**	HTGD = -1; Total_A_P <= 53.0
7	Liverpool vs. ManchesterUnited	**1**	**49.7%**	HTGD = 0; FGS = 2
8	NewcastleUnited vs. NorwichCity	**2**	**73.5%**	HTGD = 1; Total_H_P <= 67.0
9	ManchesterCity vs. Arsenal	**2**	**91.7%**	HTGD = 1; Total_H_P > 67.0
10	TottenhamHotspur vs. QueensParkRangers	**0**	**46.1%**	HTGD = -1; Total_A_P <= 53.0
11	Arsenal vs. Chelsea	**0**	**49.1%**	HTGD = 0; FGS = 0
12	Everton vs. Southampton	**2**	**93.1%**	HTGD = 2
13	Fulham vs. ManchesterCity	**1**	**39.8%**	HTGD = 0; FGS = 1
14	NorwichCity vs. Liverpool	**0**	**91.3%**	HTGD = -2
15	Reading vs. NewcastleUnited	**1**	**49.7%**	HTGD = 0; FGS = 2
16	StokeCity vs. SwanseaCity	**2**	**93.1%**	HTGD = 2
17	Sunderland vs. WiganAthletic	**1**	**49.7%**	HTGD = 0; FGS = 2
18	ManchesterUnited vs. TottenhamHotspurs	**0**	**91.3%**	HTGD = -2
19	AstonVilla vs. WestBromwichAlbion	**1**	**49.7%**	HTGD = 0; FGS = 2
20	QueensParkRangers vs. WestHam	**0**	**91.3%**	HTGD = -2

Here, we can see that there are **2 matches where the confidence of the predicted outcome is the highest, namely Match 12 and Match 16**. Both these Matches predict that the home team will win with 93.1% confidence. So only considering this model Peter can make a bet on either Match 12 or Match 16, by betting home team in either case.

Now by combining the results from both Model, Match 12 seems to be the best Match to bet on (Home team) since both the model predicted the outcome for it with the highest confidence/probability.

Match 16 is a close second, even though CHAID provided the same confidence for both matches, the multinomial logistic regressor had a slightly less probability for the home team wining for Match 16 when compared to Match 12.