IDS 572 HW4 : A GAME OF TWO HALVES

Navya Yadagiri : 674788385
Sayali Bonawale : 656488690
Jona Siko : 651224838

1. **Write the equation that Peter can use for predicting the probability of win for the home team(coded as 2) using statistically significant variables (use α = 0.05).**

As the case study provides, the target variable levels are as follows:

- 0: Loss
- 1: Draw
- 2: Win

The case study also states that Loss = 0 is the reference class,

Win equation:

Log(P(Win)/P(Loss)) = 3.313+.0035(Points_H)-0.035(Points_A)+1.618(HTGD)+0.010(Total_H_P)-0.015(Total_A_P)-3.32([FGS=0])-2.473 ([FGS=1])

Log (P(Draw)/P(Loss)) = 3.535 + 0.024(Points_H)-0.018(POINTS_A)+0.511(HTGD)-0.01(Total_A_P)-3.521([FGS=0])-2.819([FGS=bi1])

Formula:

P(Win) = exp^(Log(P(Win)/P(Loss))) / (1 + exp^(Log(P(Win)/P(Loss))) + exp^(Log(P(Draw)/P(Loss))))

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

The independent variables of RED_H(p = 0.599) and RED_A(p = 0.391) have p-values greater than 5%. The regression coefficients of this variable is not statistically significant. It can therefore be assumed that these independent variables will not influence the probability of guessing on Draw match output of red cards conceded by the home and away team.

The independent variables of RED_H(p = 0.275) and RED_A(p = 0.072) have p-values greater than 5%. The regression coefficients of this variable is not statistically significant. It can therefore be assumed that these independent variables will not influence the probability of guessing on Win match output of red cards conceded by the home and away team.

As a result, whether the away or home team receives a red card, it has no impact on the outcome of the game.

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

The independent variables of TOTAL_H_P (0.960) have p-values greater than 5%. The regression coefficients of this variable are not statistically significant. It can therefore be assumed that it is not relevant to use these points scored by the team in the previous season since it will not influence the probability of predicting the DRAW outcome of a match. Whereas, the independent variables of TOTAL_A_P (p = 0.007) have p-values less than 5%. The regression coefficients of this variable are statistically significant. It can therefore be assumed that it is not relevant to use these points scored by the team in the previous season since it will influence the probability of predicting the DRAW outcome of a match.

The independent variables of TOTAL_H_P (0.007) and TOTAL_A_P (0.000) have p-values less than 5%. The regression coefficients of these variables are statistically significant. It can therefore be assumed that it is relevant to use these points scored by the team in the previous season since it will influence the probability of predicting the WIN outcome of a match.

4. **What is the probability that the home team will win the match for the values shown in the following table?**

Using the equations below:

Log(P(Win)/P(Loss))=3.313+.0035(Points_H)-0.035(Points_A)+1.618(HTGD)+0.010(Total_H_P)-0.015(Total_A_P) -3.32([FGS=0]) -2.473 ([FGS=1])

Log (P(Draw)/P(Loss)) = 3.535 + 0.024(Points_H)-0.018(POINTS_A)+0.511(HTGD)-0.01(Total_A_P)-3.521([FGS=0])-2.819([FGS=1])

Draw: 3.535 + 0.024(15) - 0.018(18) + 0.511(2) -0.010(30) - 3.521(0) - 2.819(1)= 1.474

Win: 3.313+0.035(15)-0.035(18)+1.618(2)+0.010(40)+-0.015(30)-3.32(0)-2.473(1)= 3.921

Using the formula: P(Win) = exp^(Log(P(Win)/P(Loss))) / (1 + exp^(Log(P(Win)/P(Loss))) + exp^(Log(P(Draw)/P(Loss))))

P(Win) = exp^(3.921)/(1+exp^(3.921)+exp(1.474)) = 0 .9039

The probability of the home team winning is 90.39%.

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.**

   If the first goal is scored by the away team, we should calculate the probability f winning of the away team, using the equation below:

   Log(P(Win)/P(Loss))=3.313+.0035(Points_H)-0.035(Points_A)+1.618(HTGD)+0.010(Total_H_P)-0.015(Total_A_P) -3.32([FGS=0]) -2.473 ([FGS=1])

   Log(P(Win)/P(Loss)) = 3.313 + 0.0035(1)-0.035(1)+1.618(1)+0.010(1)-0.015(1)-3.320(1) = 1.606

   So, now the probability of Away Team winning after the first goal has been made,

   P(Win) = e^1.606 / (1 + e ^ 1.606 + e^1.474) = 0.533

   So now we can conclude that the away team has a probability of winning after the first goal is made is around 53%. We can also state that the team can be advised to bet in favor of winning since it has a greater chance of winning, since the probability is higher.

6. **What conclusions can be derived from the classification table shown in Exhibit 8? Is it advisable to bet on draws (based on the model developed)?**

   Out of 3 observations, we notice that two observations have a high correct percentage, 79.8% and 78.9% respectively, but one is quite low at 25.5%. The overall percentage of predicting correctly is 64.9%. Since one observation is quite low, we would advise to not guess on draws since it is possible that you may predict correctly is only 25.5%.

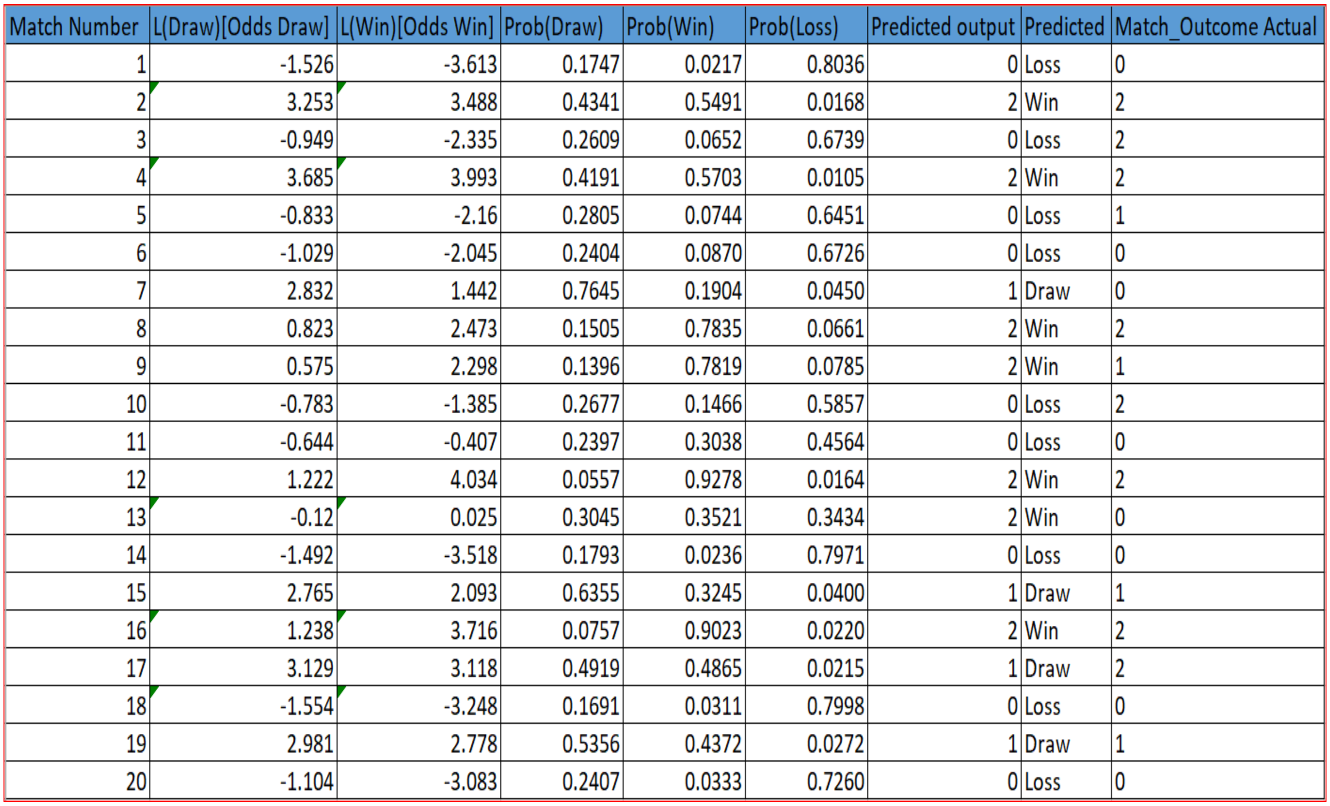
7. **Using the CHAID decision tree shown in Exhibit 9, frame rules that can be used for betting.**

   Using Exhibit 9, we can say that the root node is HTGD, since it has the highest chi-square value, and the following are the 9 decision rules that we can frame looking at the decision tree. The class that has the highest probability is assigned the target variable.

Terminal Node	Decision Rule	Support	Confidence
Node 7	If HTGD=0 & FGS =1 => Match_0 = 1	93/1520 = 0.061	37/93 = 0.397 39%
Node 6	If HTGD=0 & FGS =2 => Match_0 = 1	314/1520 0.2065	156/314 = 0.4968 49%
Node 8	If HTGD=0 & FGS =0 => Match_0 = 0	224/1520 = 0.147	110/224=0.49 49%
Node 2	If HTGD=2,3,4 => Match_0 = 2	173/1520 = 0.1138	161/173=0.93

			93%
Node 9	If HTGD=1 & Total_H_P <=67.0 => Match_0 = 2	294/1520=0.193	216/294=0.7346 73%
Node 10	If HTGD=1 & Total_H_P > 67.0 => Match_0 = 2	84/1520=0.055	77/84=0.916 =92%
Node 11	If HTGD = -1,-4 & Total_A_P <= 53.0 => Match_0 = 0	128/1520=0.0842	59/128=0.46 = 46%
Node 12	If HTGD = -1,-4 & Total_A_P > 53.0 => Match_0 = 0	141/1520=0.092	100/141=0.709 70.9%
Node 5	If HTGD = -2,-3,-5 & => Match_0 = 0	69/1520=0.045	63/69=0.913 91%

8. **Exhibit 10 lists 20 matches played over two weekends in 2012 along with the values of covariates. Use multinomial logistic regression to predict the match outcome in all 20 cases listed in Exhibit 10.**



Match Number	L(Draw)[Odds Draw]	L(Win)[Odds Win]	Prob(Draw)	Prob(Win)	Prob(Loss)	Predicted output	Predicted	Match_Outcome Actual
1	-1.526	-3.613	0.1747	0.0217	0.8036	0	Loss	0
2	3.253	3.488	0.4341	0.5491	0.0168	2	Win	2
3	-0.949	-2.335	0.2609	0.0652	0.6739	0	Loss	2
4	3.685	3.993	0.4191	0.5703	0.0105	2	Win	2
5	-0.833	-2.16	0.2805	0.0744	0.6451	0	Loss	1
6	-1.029	-2.045	0.2404	0.0870	0.6726	0	Loss	0
7	2.832	1.442	0.7645	0.1904	0.0450	1	Draw	0
8	0.823	2.473	0.1505	0.7835	0.0661	2	Win	2
9	0.575	2.298	0.1396	0.7819	0.0785	2	Win	1
10	-0.783	-1.385	0.2677	0.1466	0.5857	0	Loss	2
11	-0.644	-0.407	0.2397	0.3038	0.4564	0	Loss	0
12	1.222	4.034	0.0557	0.9278	0.0164	2	Win	2
13	-0.12	0.025	0.3045	0.3521	0.3434	2	Win	0
14	-1.492	-3.518	0.1793	0.0236	0.7971	0	Loss	0
15	2.765	2.093	0.6355	0.3245	0.0400	1	Draw	1
16	1.238	3.716	0.0757	0.9023	0.0220	2	Win	2
17	3.129	3.118	0.4919	0.4865	0.0215	1	Draw	2
18	-1.554	-3.248	0.1691	0.0311	0.7998	0	Loss	0
19	2.981	2.778	0.5356	0.4372	0.0272	1	Draw	1
20	-1.104	-3.083	0.2407	0.0333	0.7260	0	Loss	0

For all the 20 matches shown in Exhibit 10, we have listed out all the odds of draw and win, and probabilities of draw, win and loss all the levels of the target variable), and formulated the equations shown below using excel functions.

Formula:

Log(P(Win)/P(Loss)) [Odds Win] = 3.313+.0035(Points_H)-0.035(Points_A) +1.618(HTGD)+0.010(Total_H_P)-0.015(Total_A_P) -3.32([FGS=0]) -2.473 ([FGS=1])+0([FGS=2])

Log(P(Draw)/P(Loss)) [Odds Draw] = 3.535 + 0.024(Points_H)-0.018(POINTS_A) +0.511(HTGD)-0.01(Total_A_P)-3.521([FGS=0])-2.819([FGS=1])+0([FGS=2])

P(Win) = exp^(Log(P(Win)/P(Loss))) / (1 + exp^(Log(P(Win)/P(Loss))) + exp^(Log(P(Draw)/P(Loss))))

P(Draw) = exp^(Log(P(Draw)/P(Loss))) / (1 + exp^(Log(P(Win)/P(Loss))) + exp^(Log(P(Draw)/P(Loss))))

P(Win) = 1 / (1 + exp^(Log(P(Win)/P(Loss))) + exp^(Log(P(Draw)/P(Loss))))

The Predicted output we have calculated by comparing the highest probability amongst Win, Draw, Loss. We have 7 misclassified data points.

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.**

Using the CHAID Decision Tree, multinomial logistic regression below are the actual and predicted target variables values. Below we are comparing predicted values of CHAID decision tree and multinomial logistic regression.

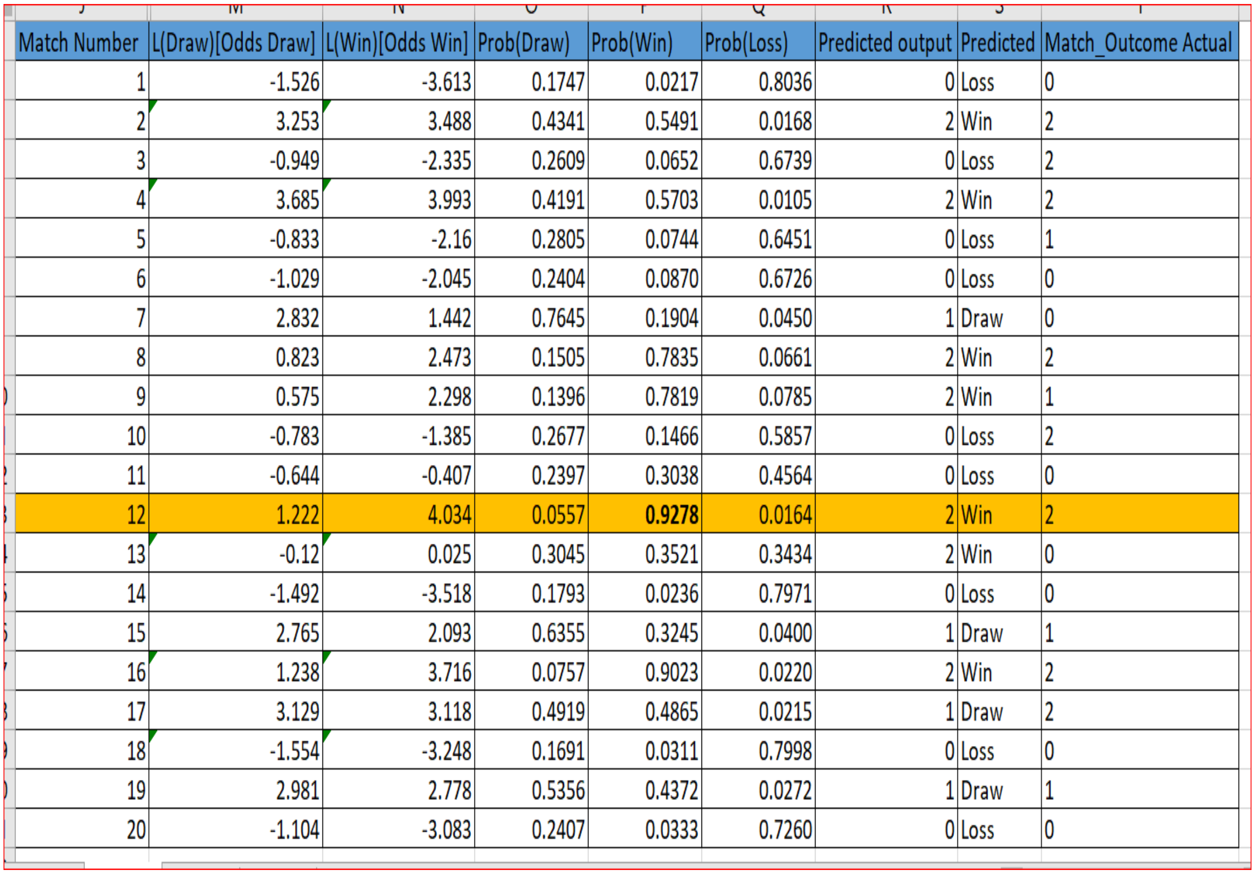
		Using CHAID Decision tree	**Using Multinomial Logistic regression**	**Comparison**
Match Number	**Match Outcome Actual**	**Predicted**	**Predicted**	**Classified correctly / Misclassified**
1	0	0	0	Classified correctly
2	2	1	2	Misclassified
3	2	0	0	Classified correctly
4	2	1	2	Misclassified
5	1	0	0	Classified correctly
6	0	0	0	Classified correctly

7	0	1	1	Classified correctly
8	2	2	2	Classified correctly
9	1	2	2	Classified correctly
10	2	0	0	Classified correctly
11	0	0	0	Classified correctly
12	2	2	2	Classified correctly
13	0	1	2	Misclassified
14	0	0	0	Classified correctly
15	1	1	1	Classified correctly
16	2	2	2	Classified correctly
17	2	1	1	Classified correctly
18	0	0	0	Classified correctly
19	1	1	1	Classified correctly

20	0	0	0	Classified correctly

Using CHAID decision tree, the Accuracy = 11/20 (True Positive + True Negative) /Total number of data points = 55%, while using the multinomial logistic regression, we have predicted 13/20 data points correctly, so we can say that the accuracy of multinomial logistic regression is 65%. When we compare both the models, we see that 3 data points have shown discrepancies (Show different target variable). Hence, the misclassification rate is 3/20 = 15% (i.e., Discrepancy between CHAID decision tree and multinomial logistic regression).

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.



Match Number	L(Draw)[Odds Draw]	L(Win)[Odds Win]	Prob(Draw)	Prob(Win)	Prob(Loss)	Predicted output	Predicted	Match_Outcome Actual
1	-1.526	-3.613	0.1747	0.0217	0.8036	0	Loss	0
2	3.253	3.488	0.4341	0.5491	0.0168	2	Win	2
3	-0.949	-2.335	0.2609	0.0652	0.6739	0	Loss	2
4	3.685	3.993	0.4191	0.5703	0.0105	2	Win	2
5	-0.833	-2.16	0.2805	0.0744	0.6451	0	Loss	1
6	-1.029	-2.045	0.2404	0.0870	0.6726	0	Loss	0
7	2.832	1.442	0.7645	0.1904	0.0450	1	Draw	0
8	0.823	2.473	0.1505	0.7835	0.0661	2	Win	2
9	0.575	2.298	0.1396	0.7819	0.0785	2	Win	1
10	-0.783	-1.385	0.2677	0.1466	0.5857	0	Loss	2
11	-0.644	-0.407	0.2397	0.3038	0.4564	0	Loss	0
12	1.222	4.034	0.0557	**0.9278**	0.0164	2	Win	2
13	-0.12	0.025	0.3045	0.3521	0.3434	2	Win	0
14	-1.492	-3.518	0.1793	0.0236	0.7971	0	Loss	0
15	2.765	2.093	0.6355	0.3245	0.0400	1	Draw	1
16	1.238	3.716	0.0757	0.9023	0.0220	2	Win	2
17	3.129	3.118	0.4919	0.4865	0.0215	1	Draw	2
18	-1.554	-3.248	0.1691	0.0311	0.7998	0	Loss	0
19	2.981	2.778	0.5356	0.4372	0.0272	1	Draw	1
20	-1.104	-3.083	0.2407	0.0333	0.7260	0	Loss	0

Using the solution to Question 8, we can conclude that Match 12 Everton vs. Southampton is the best match to bet over because it has the highest probability of winning of 92.78%, as indicated above.