Meghashree Maddihally Nagoji

4/5/2022

Context

In the provided case, historical data for football matches is provided to predict future match results which will be utilized for in-game betting.

This will be achieved using logistic regression, as there are 3 classes (Loss - 0, Draw - 1, Win - 2) to determine the result, this classifies as a multi-class classification problem.

Given: The loss, i.e.,0, will be considered as the reference variable.

Log(Home Team Win/Home Team Loss) = B_0 + Σ(i= 1 to t) B_TX

Log(Home Team Draw/Home Team Loss) = B_0 + Σ(i= 1 to t) B_TX

Probability(Win) = e^(win) / 1 + e^win + e^draw

Probability(draw) = e^(draw) / 1 + e^draw + e^win

Probability(Loss) = 1 – Probability(win) - Probability(draw)

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 alpha = 0.05).**

   To find the equation for predicting the probability of win, we need to first calculate the odds ratio of draw and win.

   After considering only the statistically significant variables (i.e., p-value<0.05) from Exhibit 7, we get the below log-odds equations:

   L(Draw) = log(Draw/Loss) = 3.535 + (0.024 * POINTS_H) – (0.018 * POINTS_A) + (0.511 * HTGD) – (0.010 * TOTAL_A_P) – (3.521 * [FGS=0]) – (2.819 * [FGS=1])

   L(Win) = log(Win/Loss) = 3.313 + (0.035 * POINTS_H) – (0.035 * POINTS_A) + (1.618 * HTGD) + (0.010 * TOTAL_H_P) – (0.015 * TOTAL_A_P) – (3.320 * [FGS=0]) – (2.473 * [FGS=1])

Peter can use the below equation for predicting the probability of win for the home team:

P(Win) = e^L(Win) / (1 + e^L(Win) + e^L(Draw))

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.**

   From Exhibit 7,
   For code 2(Win), the p-value for RED_H(*red cards for home team*) & RED_A(*red cards for away team)* is **0.275 & 0.072** respectively.
   And For code 1(Draw), the p-value for RED_H(*red cards for home team*) & RED_A(*red cards for away team*) is **0.599 & 0.391** respectively. In Both the cases, the p-value is greater than the level of significance(**alpha = 0.05**), that makes the variables statistically insignificant. They have no influence on the match output.

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

   From Exhibit 7,
   For code 2(Win), the p-value for TOTAL_H_P(*total points earned by the home team in the previous season*) & TOTAL_A_P(*total points earned by the away team in the previous season*) is **0.007 & 0.000** respectively. The p-value is less than the level of significance(**alpha = 0.05**), that makes the variables statistically significant. Therefore, it is relevant to use the points scored by both home team and away team in the previous season for predicting the outcome of a match.

   And For code 1(Draw), the p-value for TOTAL_H_P(*total points earned by the home team in the previous season*) & TOTAL_A_P(*total points earned by the away team in the previous season*) is **0.960 & 0.007** respectively. The p-value of TOTAL_H_P is more than level of significance(**alpha = 0.05**), that makes this variable statistically insignificant and does not affect the match output, however the p-value of TOTAL_A_P is less than the level of significance(**alpha = 0.05**), that makes this variable statistically significant. Therefore, it is relevant to use the points scored by only away team in the previous season for predicting the 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?**

   L(Draw) = log(Draw/Loss) = 3.535 + (0.024 * POINTS_H) – (0.018 * POINTS_A) +

(0.511 * HTGD) – (0.010 * TOTAL_A_P) – (3.521 * [FGS=0]) – (2.819 * [FGS=1])

L(Win) = log(Win/Loss) = 3.313 + (0.035 * POINTS_H) – (0.035 * POINTS_A) + (1.618 * HTGD) + (0.010 * TOTAL_H_P) – (0.015 * TOTAL_A_P) – (3.320 * [FGS=0]) – (2.473 * [FGS=1])

Substituting the values from the table given in the question in L(Draw) and L(Win):

L(Draw) = 3.535 + (0.024*15) – (0.018*18) + (0.511*2) – (0.010*30) – (3.521*0) – (2.819*1)
= **1.474**

L(Win) = 3.313 + (0.035*15) – (0.035*18) + (1.618*2) + (0.010*40) – (0.015*30) – (3.320*0) – (2.473*1)
= **3.921**

Substituting the values of L(Draw) and L(Win) in the below probability equation:
P(Win) = e^L(Win) / (1 + e^L(Win) + e^L(Draw))
= e^ (3.921) / 1 + e^ (3.921) + e^ (1.474)
= **0.9039**

Hence, the probability of 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.**

   Since, the first goal is scored by the Away team (FGS=0, i.e., *first goal scored by away team*) variable will contain the value of 1 and (FGS=1, *first goal scored by home team*) variable will contain the value of 0. Keeping the rest of the variables as is from the question 4.

   L(Draw) = log(Draw/Loss) = 3.535 + (0.024 * POINTS_H) – (0.018 * POINTS_A) + (0.511 * HTGD) – (0.010 * TOTAL_A_P) – (3.521 * [FGS=0]) – (2.819 * [FGS=1])

   L(Win) = log(Win/Loss) = 3.313 + (0.035 * POINTS_H) – (0.035 * POINTS_A) + (1.618 * HTGD) + (0.010 * TOTAL_H_P) – (0.015 * TOTAL_A_P) – (3.320 * [FGS=0]) – (2.473 * [FGS=1])

Substituting FGS=0 as 1 and FGS=1 as 0 in L(Draw) and L(Win):

L(Draw) = 3.535 + (0.024*15) – (0.018*18) + (0.511*2) – (0.010*30) – (3.521*1) – (2.819*0)
= **0.772**

L(Win) = 3.313 + (0.035*15) – (0.035*18) + (1.618*2) + (0.010*40) – (0.015*30) – (3.320*1) – (2.473*0)
= **3.074**

Substituting the values of L(Draw) and L(Win) in the below probability equations:

P(Draw) = e^L(Draw) / (1 + e^L(Win) + e^L(Draw))
P(Draw) = e^(0.772) / (1 + e^(3.074) + e^(0.772))
= **0.0872**

P(Win) = e^L(Win) / (1 + e^L(Win) + e^L(Draw))
P(Win) = e^(3.074) / (1 + e^(3.074) + e^(0.772))
= **0.8723**

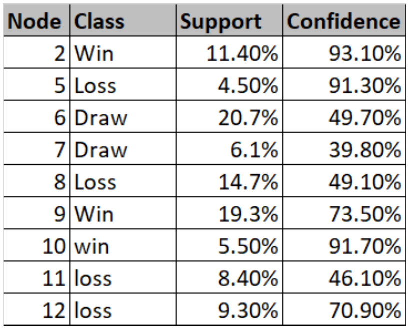
We know that the probability of P(Loss) + P(Draw) + P(Win) =1
Therefore, P(Loss) = 1 – P(Win) – P(Draw)
= 1 – 0.8723 – 0.0872
= 0.0405
P(Loss) = 4.05%

As the probability of losing by the home team when the first goal is scored by the away team is only 4.05%, it is not advisable to bet in favor of the away team.

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

   From Exhibit 8, we can observe that the correct percentage for draw i.e., saying observed draw is predicted correctly as draw is only 103 out of 404 (120+103+181). So, it is **not advisable** to bet on draws. Since, only **25.5 %** of bets will be successful.

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



Node	Class	Support	Confidence
2	Win	11.40%	93.10%
5	Loss	4.50%	91.30%
6	Draw	20.7%	49.70%
7	Draw	6.1%	39.80%
8	Loss	14.7%	49.10%
9	Win	19.3%	73.50%
10	win	5.50%	91.70%
11	loss	8.40%	46.10%
12	loss	9.30%	70.90%

From exhibit 9 in the provided case, we have the above-mentioned support (calculated) and confidence.

Strong winning bet rules depending on the match outcome:

Win

a. If the half time goal difference(HTGD) is 2,3 or 4 then the home team will win the match.
   Support 11.4% Confidence 93.1%
b. If the half time goal difference(HTGD) is 1 and total points earned by the home team in the previous season(TOTAL_H_P) is greater than 67 then the home team will win the match.
   Support 5.5% Confidence 91.7%

Loss

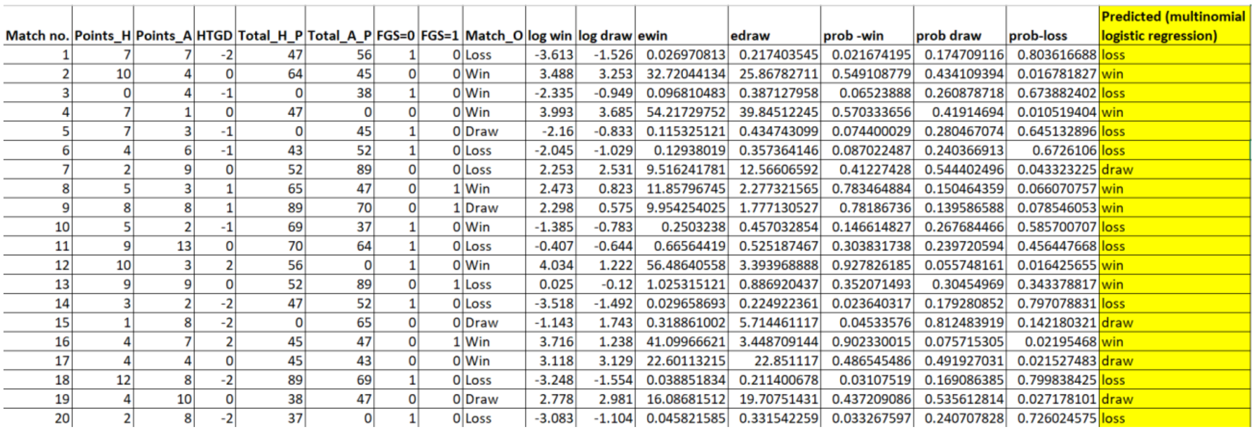
a. If the half time goal difference(HTGD) is -2,-3 or -5 then the home team will lose the match.
   Support 4.5%, Confidence 91.3%
b. If the half time goal difference is -1 or -4 and total points earned by the away team in previous season(TOTAL_A_P) is greater than 53 points, then the home team will lose the match.
   Support 9.3%, Confidence 70.9%

Calculated Bets:

a. If the half time goal difference(HTGD) is 0 and no goals are scored (FGS = 2) then the home team will have approximately 50% chance of winning or ending the match in a draw, therefore a bet can be placed in these two results to minimize the loss.
   Support 20.7%, Confidence(draw) 49.7% Confidence(win) 48.1%

8. Exhibit 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.

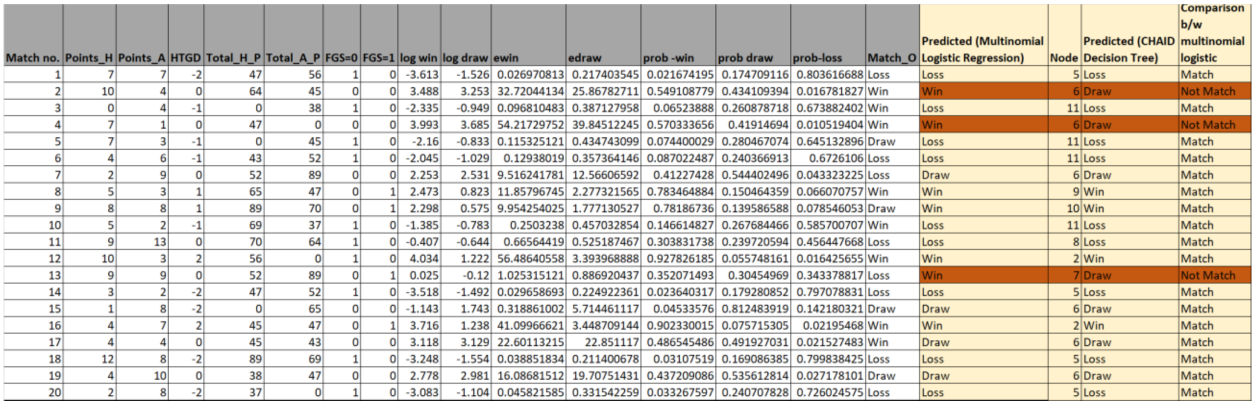
The below table shows the predictions of the match outcome in all 20 cases listed in Exhibit 10.



Match no.	Points_H	Points_A	HTGD	Total_H_P	Total_A_P	FGS=0	FGS=1	Match_O	log win	log draw	ewin	edraw	prob -win	prob draw	prob-loss	Predicted (multinomial logistic regression)
1	7	7	-2	47	56	1	0	Loss	-3.613	-1.526	0.026970813	0.217403545	0.021674195	0.174709116	0.803616688	loss
2	10	4	0	64	45	0	0	Win	3.488	3.253	32.72044134	25.86782711	0.549108779	0.434109394	0.016781827	win
3	0	4	-1	0	38	1	0	Win	-2.335	-0.949	0.096810483	0.387127958	0.06523888	0.260878718	0.673882402	loss
4	7	1	0	47	0	0	0	Win	3.993	3.685	54.21729752	39.84512245	0.570333656	0.41914694	0.010519404	win
5	7	3	-1	0	45	1	0	Draw	-2.16	-0.833	0.115325121	0.434743099	0.074400029	0.280467074	0.645132896	loss
6	4	6	-1	43	52	1	0	Loss	-2.045	-1.029	0.12938019	0.357364146	0.087022487	0.240366913	0.6726106	loss
7	2	9	0	52	89	0	0	Loss	2.253	2.531	9.516241781	12.56606592	0.41227428	0.544402496	0.043323225	draw
8	5	3	1	65	47	0	1	Win	2.473	0.823	11.85796745	2.277321565	0.783464884	0.150464359	0.066070757	win
9	8	8	1	89	70	0	1	Draw	2.298	0.575	9.954254025	1.777130527	0.78186736	0.139586588	0.078546053	win
10	5	2	-1	69	37	1	0	Win	-1.385	-0.783	0.2503238	0.457032854	0.146614827	0.267684466	0.585700707	loss
11	9	13	0	70	64	1	0	Loss	-0.407	-0.644	0.66564419	0.525187467	0.303831738	0.239720594	0.456447668	loss
12	10	3	2	56	0	1	0	Win	4.034	1.222	56.48640558	3.393968888	0.927826185	0.055748161	0.016425655	win
13	9	9	0	52	89	0	1	Loss	0.025	-0.12	1.025315121	0.886920437	0.352071493	0.30454969	0.343378817	win
14	3	2	-2	47	52	1	0	Loss	-3.518	-1.492	0.029658693	0.224922361	0.023640317	0.179280852	0.797078831	loss
15	1	8	-2	0	65	0	0	Draw	-1.143	1.743	0.318861002	5.714461117	0.04533576	0.812483919	0.142180321	draw
16	4	7	2	45	47	0	1	Win	3.716	1.238	41.09966621	3.448709144	0.902330015	0.075715305	0.02195468	win
17	4	4	0	45	43	0	0	Win	3.118	3.129	22.60113215	22.851117	0.486545486	0.491927031	0.021527483	draw
18	12	8	-2	89	69	1	0	Loss	-3.248	-1.554	0.038851834	0.211400678	0.03107519	0.169086385	0.799838425	loss
19	4	10	0	38	47	0	0	Draw	2.778	2.981	16.08681512	19.70751431	0.437209086	0.535612814	0.027178101	draw
20	2	8	-2	37	0	1	0	Loss	-3.083	-1.104	0.045821585	0.331542259	0.033267597	0.240707828	0.726024575	loss

The calculations for the above table are present in the excel file
'HW4_Megha_Roopesh_Skandan.xlsx', sheet - question 8.

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

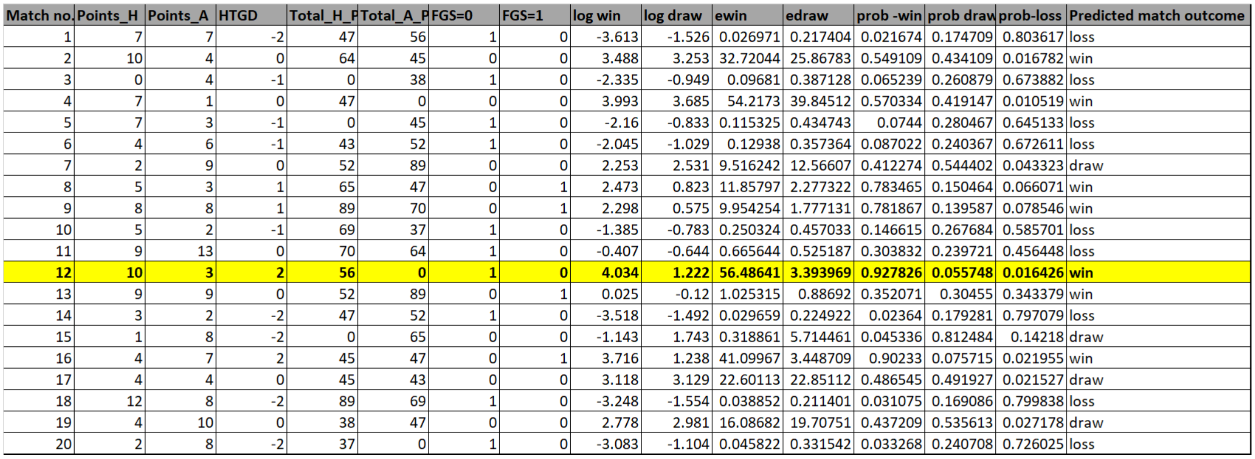


Match no.	Points_H	Points_A	HTGD	Total_H_P	Total_A_P	FGS=0	FGS=1	log win	log draw	ewin	edraw	prob -win	prob draw	prob-loss	Match_O	Predicted (Multinomial Logistic Regression)	Node	Predicted (CHAID Decision Tree)	Comparison b/w multinomial logistic
1	7	7	-2	47	56	1	0	-3.613	-1.526	0.026970813	0.217403545	0.021674195	0.174709116	0.803616688	Loss	Loss	5	Loss	Match
2	10	4	0	64	45	0	0	3.488	3.253	32.72044134	25.86782711	0.549108779	0.434109394	0.016781827	Win	Win	6	Draw	Not Match
3	0	4	-1	0	38	1	0	-2.335	-0.949	0.096810483	0.387127958	0.06523888	0.260878718	0.673882402	Win	Loss	11	Loss	Match
4	7	1	0	47	0	0	0	3.993	3.685	54.21729752	39.84512245	0.570333656	0.41914694	0.010519404	Win	Win	6	Draw	Not Match
5	7	3	-1	0	45	1	0	-2.16	-0.833	0.115325121	0.434743099	0.074400029	0.280467074	0.645132896	Draw	Loss	11	Loss	Match
6	4	6	-1	43	52	1	0	-2.045	-1.029	0.12938019	0.357364146	0.087022487	0.240366913	0.6726106	Loss	Loss	11	Loss	Match
7	2	9	0	52	89	0	0	2.253	2.531	9.516241781	12.56606592	0.41227428	0.544402496	0.043323225	Loss	Draw	6	Draw	Match
8	5	3	1	65	47	0	1	2.473	0.823	11.85796745	2.277321565	0.783464884	0.150464359	0.066070757	Win	Win	9	Win	Match
9	8	8	1	89	70	0	1	2.298	0.575	9.954254025	1.777130527	0.78186736	0.139586588	0.078546053	Draw	Win	10	Win	Match
10	5	2	-1	69	37	1	0	-1.385	-0.783	0.2503238	0.457032854	0.146614827	0.267684466	0.585700707	Win	Loss	11	Loss	Match
11	9	13	0	70	64	1	0	-0.407	-0.644	0.66564419	0.525187467	0.303831738	0.239720594	0.456447668	Loss	Loss	8	Loss	Match
12	10	3	2	56	0	1	0	4.034	1.222	56.48640558	3.393968888	0.927826185	0.055748161	0.016425655	Win	Win	2	Win	Match
13	9	9	0	52	89	0	1	0.025	-0.12	1.025315121	0.886920437	0.352071493	0.30454969	0.343378817	Loss	Win	7	Draw	Not Match
14	3	2	-2	47	52	1	0	-3.518	-1.492	0.029658693	0.224922361	0.023640317	0.179280852	0.797078831	Loss	Loss	5	Loss	Match
15	1	8	-2	0	65	0	0	-1.143	1.743	0.318861002	5.714461117	0.04533576	0.812483919	0.142180321	Draw	Draw	6	Draw	Match
16	4	7	2	45	47	0	1	3.716	1.238	41.09966621	3.448709144	0.902330015	0.075715305	0.02195468	Win	Win	2	Win	Match
17	4	4	0	45	43	0	0	3.118	3.129	22.60113215	22.851117	0.486545486	0.491927031	0.021527483	Win	Draw	6	Draw	Match
18	12	8	-2	89	69	1	0	-3.248	-1.554	0.038851834	0.211400678	0.03107519	0.169086385	0.799838425	Loss	Loss	5	Loss	Match
19	4	10	0	38	47	0	0	2.778	2.981	16.08681512	19.70751431	0.437209086	0.535612814	0.027178101	Draw	Draw	6	Draw	Match
20	2	8	-2	37	0	1	0	-3.083	-1.104	0.045821585	0.331542259	0.033267597	0.240707828	0.726024575	Loss	Loss	5	Loss	Match

As per the worksheet mentioned above, among the 20 matches played, multinomial logistic regression has been able to predict 17 (marked in light gold) out of 20 predictions correctly in comparison to the CHAID decision tree. The incorrect predictions are marked in brown.

The calculations for the above table are present in the excel file
'HW4_Megha_Roopesh_Skandan.xlsx', sheet - question 9.

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 no.	Points_H	Points_A	HTGD	Total_H_P	Total_A_P	FGS=0	FGS=1	log win	log draw	ewin	edraw	prob -win	prob draw	prob-loss	Predicted match outcome
1	7	7	-2	47	56	1	0	-3.613	-1.526	0.026971	0.217404	0.021674	0.174709	0.803617	loss
2	10	4	0	64	45	0	0	3.488	3.253	32.72044	25.86783	0.549109	0.434109	0.016782	win
3	0	4	-1	0	38	1	0	-2.335	-0.949	0.09681	0.387128	0.065239	0.260879	0.673882	loss
4	7	1	0	47	0	0	0	3.993	3.685	54.2173	39.84512	0.570334	0.419147	0.010519	win
5	7	3	-1	0	45	1	0	-2.16	-0.833	0.115325	0.434743	0.0744	0.280467	0.645133	loss
6	4	6	-1	43	52	1	0	-2.045	-1.029	0.12938	0.357364	0.087022	0.240367	0.672611	loss
7	2	9	0	52	89	0	0	2.253	2.531	9.516242	12.56607	0.412274	0.544402	0.043323	draw
8	5	3	1	65	47	0	1	2.473	0.823	11.85797	2.277322	0.783465	0.150464	0.066071	win
9	8	8	1	89	70	0	1	2.298	0.575	9.954254	1.777131	0.781867	0.139587	0.078546	win
10	5	2	-1	69	37	1	0	-1.385	-0.783	0.250324	0.457033	0.146615	0.267684	0.585701	loss
11	9	13	0	70	64	1	0	-0.407	-0.644	0.665644	0.525187	0.303832	0.239721	0.456448	loss
12	**10**	**3**	**2**	**56**	**0**	**1**	**0**	**4.034**	**1.222**	**56.48641**	**3.393969**	**0.927826**	**0.055748**	**0.016426**	**win**
13	9	9	0	52	89	0	1	0.025	-0.12	1.025315	0.88692	0.352071	0.30455	0.343379	win
14	3	2	-2	47	52	1	0	-3.518	-1.492	0.029659	0.224922	0.02364	0.179281	0.797079	loss
15	1	8	-2	0	65	0	0	-1.143	1.743	0.318861	5.714461	0.045336	0.812484	0.14218	draw
16	4	7	2	45	47	0	1	3.716	1.238	41.09967	3.448709	0.90233	0.075715	0.021955	win
17	4	4	0	45	43	0	0	3.118	3.129	22.60113	22.85112	0.486545	0.491927	0.021527	draw
18	12	8	-2	89	69	1	0	-3.248	-1.554	0.038852	0.211401	0.031075	0.169086	0.799838	loss
19	4	10	0	38	47	0	0	2.778	2.981	16.08682	19.70751	0.437209	0.535613	0.027178	draw
20	2	8	-2	37	0	1	0	-3.083	-1.104	0.045822	0.331542	0.033268	0.240708	0.726025	loss

As per the line item marked in yellow in the above worksheet, match 12 (Everton Vs Southampton) has the highest probability of winning : 92.78% (prob-win). The match outcome is predicted to be win from both multinomial logistic regression model and CHAID tree.