CSE 40647/60647: Data Science	**Fall 2017**
Homework 4	
Handed Out: October 12, 2017	*Due: November 09, 2017 11:59 pm*

1 General Instructions

- This assignment is due at 11:59 PM on the due date.
- We will be using Sakai (https://sakailogin.nd.edu/portal/site/FA17-CSE-40647-CX-01) for collecting this assignment. Contact TA if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The homework MUST be submitted in pdf format. You can handwrite trees/figures and scan them into PDF. Name your pdf file as YourNetid-HW4.pdf.
- Please use Piazza if you have questions about the homework. Also feel free to send TA emails and come to office hours.

2 Question 1 (20 points)

ID3 model, a decision tree model using "Information Gain"

1. Construct a decision tree based on the training set (24 games).
2. Use the decision tree to predict labels of instances in the testing set (6 games) based on their attributes.
3. Calculate Accuracy, Precision, Recall, and F-1 score on the testing set, given the ground-truth labels of the 6 games.
4. Predict labels of the future 4 games using the decision tree.

Solutions:

2.1 Construct decision tree - ID3

2.1.1 Step 1:

1. We want to predict whether a game is win or loss based on 3 attributes.
2. Y: (Win*14, Lose*10), the entropy $H(Y) = -\sum_{i=1}^{m} p_i * logp_i = -\frac{14}{24}log_2\frac{14}{24} - \frac{10}{24}log_2\frac{10}{24} = 0.9799$.
3. $X_{HomeAway}$: (Home*15, Away*9); Home: (Win*10, Lose*5), Away: (Win*4, Lose*5).

4. The conditional entropy $H(Y|X_{HomeAway}) = \sum_x p(x) * H(Y|X = x) = \frac{15}{24}(-\frac{10}{15}log_2\frac{10}{15} - \frac{5}{15}log_2\frac{5}{15}) + \frac{9}{24}(-\frac{4}{9}log_2\frac{4}{9} - \frac{5}{9}log_2\frac{5}{9}) = 0.9456$.
5. Then Information Gain is $IG(Y|X_{HomeAway}) = H(Y) - H(Y|X_{HomeAway}) = 0.9799 - 0.9456 = 0.0343$.
6. Similar as above, we can get $IG(Y|X_{AP25}) = 0.1091$ and $IG(Y|X_{Media}) = 0.1891$.
7. Then we choose the attributes with maximum Information Gain which is X_{Media}.
8. Then we will have our first tree in Figure 1.

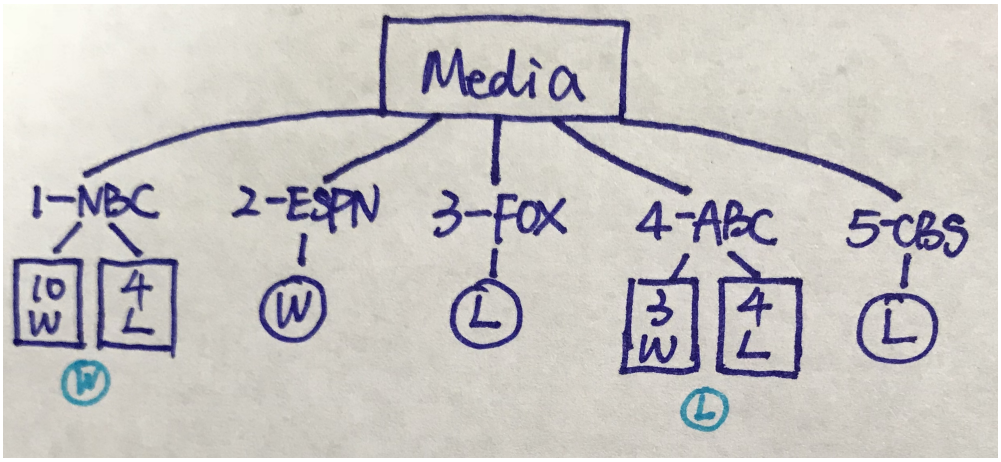

Figure 1: ID3-Tree1

2.1.2 Step 2:

1. Since we already picked Media as our first attributes, we can tell from Tree1, we need add extra attribute into our decision tree.
2. Given media at 1-NBC, we total have 14 events, Y: (Win*10, Lose*4), $H(Y|NBC) = -\frac{10}{14}log_2\frac{10}{14} - \frac{4}{14}log_2\frac{4}{14} = 0.8631$
3. $X_{HomeAway}$: (Home*13, Away*1); Home: (Win*9, Lose*4), Away: (Win*1).
4. $H(Y|X_{HomeAway}, NBC) = \frac{13}{14}(-\frac{9}{13}log_2\frac{9}{13} - \frac{4}{13}log_2\frac{4}{13}) + \frac{1}{14}(-\frac{1}{1}log_2\frac{1}{1}) = 0.8269$.
5. $IG(Y|X_{HomeAway}, NBC) = 0.8631 - 0.8269 = 0.0362$.
6. Similar as above, we can get $IG(Y|X_{AP25}, NBC) = 0.0617$.
7. Given media at 4-ABC, we total have 7 events, Y: (Win*3, Lose*4), the entropy $H(Y|ABC) = 0.9852$
8. $IG(Y|X_{HomeAway}, ABC) = 0$ and $IG(Y|X_{AP25}, ABC) = 0.2917$.

9. We choose attributes for maximum information gain, both X_{AP25} for 1-NBC and 4-ABC.

10. Then we will have our second tree in Figure 2.

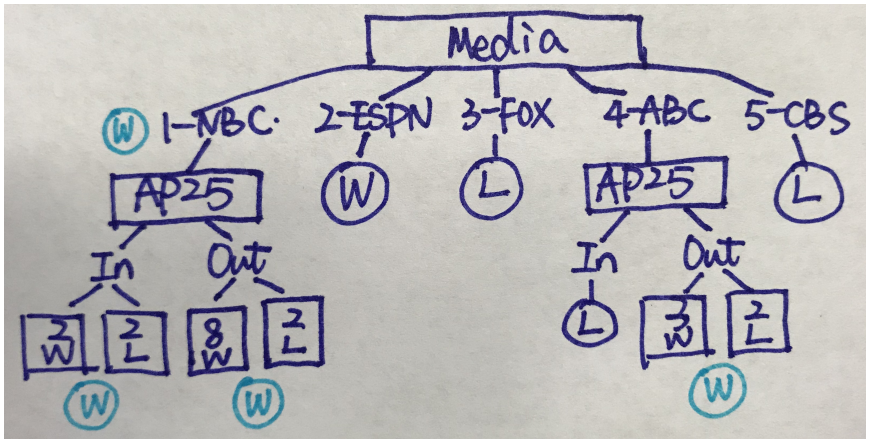

Figure 2: ID3-Tree2

2.1.3 Step 3:

1. Given media at 1-NBC, opponent in AP25, 4 events, Y: (Win*2, Lose*2), $H(Y|NBC, In) = 1$.

2. $H(Y|X_{HomeAway}, NBC, In) = 0$, no information gain.

3. Given media at 1-NBC, opponent out AP25, 10 events, Y: (Win*8, Lose*2), $H(Y|NBC, Out) = 0.7219$.

4. $H(Y|X_{HomeAway}, NBC, Out) = 0.6878$ and $IG(Y|X_{HomeAway}, NBC, Out) = 0.0341$.

5. Given Media at 4-ABC, opponent out AP25, 5 events, Y: (Win*3, Lose*2).

6. $IG(Y|X_{HomeAway}, ABC, Out) = 0$, No information gain.

7. Then we will have our final decision tree in Figure 3.

8. Note: in practice, if the requirements on output allow us to give half/half predictions (it's like three-label: Win, Loss, Not Available), then we can give an N/A; if not, the classifier will find a binary result – either take the most frequent class label in the dataset (it's like decision at the root node of the tree) or randomly give a prediction (Win or Loss). Here, we allow students to take either way; in classification evaluation (accuracy, precision, recall), the students can either drop this entry from the test set or give a binary answer. The answer here provided is based on it's root node.

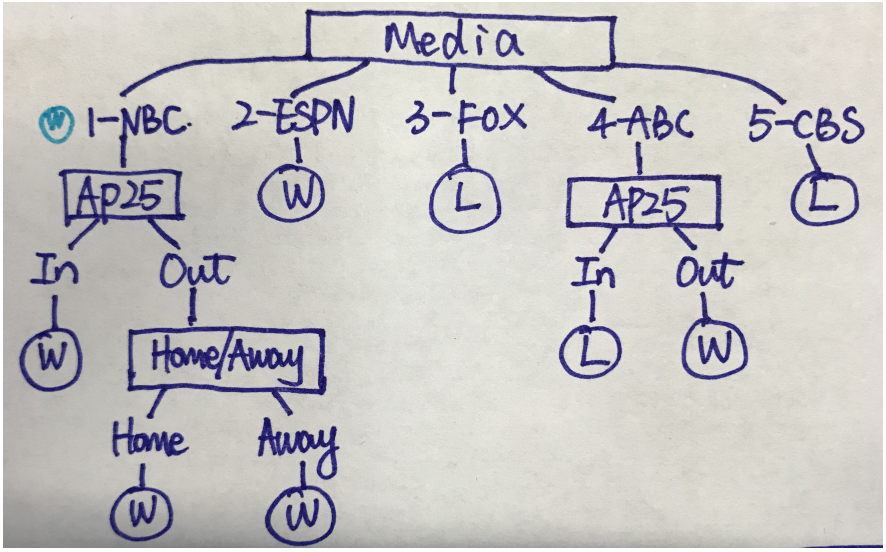

Figure 3: ID3-Final Decision Tree

2.2 Label testing sets based on the ID3 tree

ID	Media	AP25	Home/Away	Label	Truth
25	1-NBC	Out	Home	Win	Win
26	1-NBC	In	Home	Win	Lose
27	2-ESPN			Win	Win
28	3-FOX			Lose	Win
29	1-NBC	Out	Home	Win	Win
30	4-ABC	Out	Away	Win	Win

2.3 Evaluation

Actual/Prediction	Win	Lose	Total
Win	4 (TP)	1 (FN)	5
Lose	1 (FP)	0 (TN)	1
Total	5	1	6

1. $Accuracy = \frac{TP+TN}{All} = \frac{4}{6} = 0.667.$
2. $Precision = \frac{TP}{TP+FP} = \frac{4}{5} = 0.8.$
3. $Recall = \frac{TP}{TP+FN} = \frac{4}{5} = 0.8.$
4. $F_1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*0.8*0.8}{0.8+0.8} = 0.8.$

2.4 Predictions

ID	Media	AP25	Home/Away	Prediction
31	1-NBC	In		Win
32	1-NBC	Out	Home	Win
33	1-NBC	Out	Home	Win
35	1-NBC	Out	Home	Win

3 Question 2 (20 points)

C4.5 model, a decision tree model using "Gain Ratio".

1. Construct a decision tree based on the training set (24 games).
2. Use the decision tree to predict labels of instances in the testing set (6 games) based on their attributes.
3. Calculate Accuracy, Precision, Recall, and F-1 score on the testing set, given the ground-truth labels of the 6 games.
4. Predict labels of the future 4 games using the decision tree.

Solutions:

3.1 Construct decision tree - ID4.5

3.1.1 Step 1:

1. We want to predict whether a game is win or loss based on 3 attributes.
2. According to ID3 model, we have information gain: $IG(Y|X_{HomeAway}) = 0.0343$, $IG(Y|X_{AP25}) = 0.1091$ and $IG(Y|X_{Media}) = 0.1891$.
3. (Home*15, Away*9), $SplitInfo(X_{HomeAway}) = -\frac{15}{24}log_2\frac{15}{24} - \frac{9}{24}log_2\frac{9}{24} = 0.9544$.
4. $GainRatio(X_{HomeAway}) = 0.0343/0.9544 = 0.0359$.
5. Similar as above, we can get $GainRatio(X_{AP25}) = 0.1253$ and $GainRatio(X_{Media}) = 0.1224$.
6. Then we choose the attributes with maximum Gain Ratio which is X_{AP25}.
7. Then we will have our first tree in Figure 4.

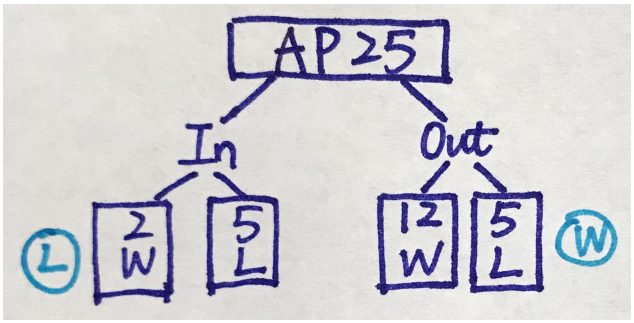

Figure 4: ID4.5-Tree1

3.1.2 Step 2:

1. Since we already picked AP25 as our first attributes, we can tell from Tree1, we need add extra attribute into our decision tree.
2. Given opponent in AP25, we total have 7 events, (Home*4, Away*3), Home: (Win*2, Lose*2), Away: (Lose*3) and $GainRatio(X_{HomeAway}) = 0.2961$.
3. (1-NBC*4, 3-FOX*1, 4-ABC*2), NBC: (Win*2, Lose*2), FOX: (Lose*1), ABC: (Lose*2) and $GainRatio(X_{Media}) = 0.2116$
4. Given opponent out AP25, we total have 17 events, (Home*11, Away*6), Home: (Win*8, Lose*3), Away: (Win*4, Lose*2) and $GainRatio(X_{HomeAway}) = 0.0031$.
5. (1-NBC*10, 2-ESPN*1, 4-ABC*5, 5-CBS*1), NBC: (Win*8, Lose*2), ESPN: (Win*1), ABC: (Win*3, Lose*2), ABC: (Lose*1) and $GainRatio(X_{Media}) = 0.1129$.
6. We choose attributes for maximum Gain Ratio, $X_{HomeAway}$ for opponent in AP25 and X_{Media} for opponent out AP25.
7. Then we will have our second tree in Figure 5.

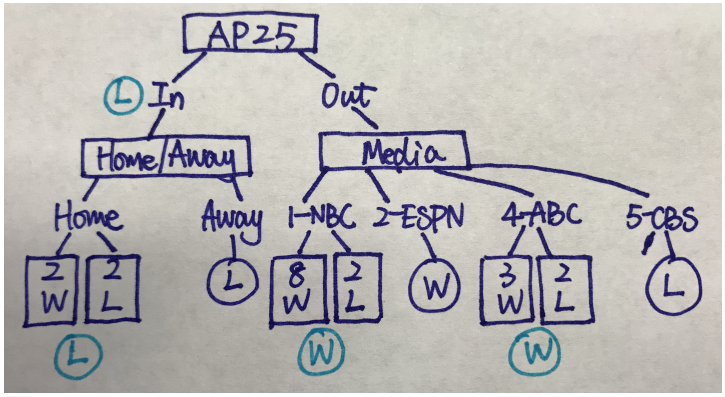

Figure 5: ID4.5-Tree2

3.1.3 Step 3:

1. Given opponent in AP25 at home, 4 events, 1-NBC: (Win*2, Lose*2), No information gain for media.
2. Given opponent out AP25 at 1-NBC, 10 events, (Home*9, Away*1), Home: (Win*7, Lose*2) and Away: (Win*1) and $GainRatio(X_{HomeAway}) = 0.0727$.
3. Given opponent out AP25 at 4-ABC, 5 events, Away: (Win*3, Lose*2), No information gain for HomeAway.
4. Then we will have our final decision tree in Figure 6.
5. The answer provided here for 50/50 is based on it's root node.

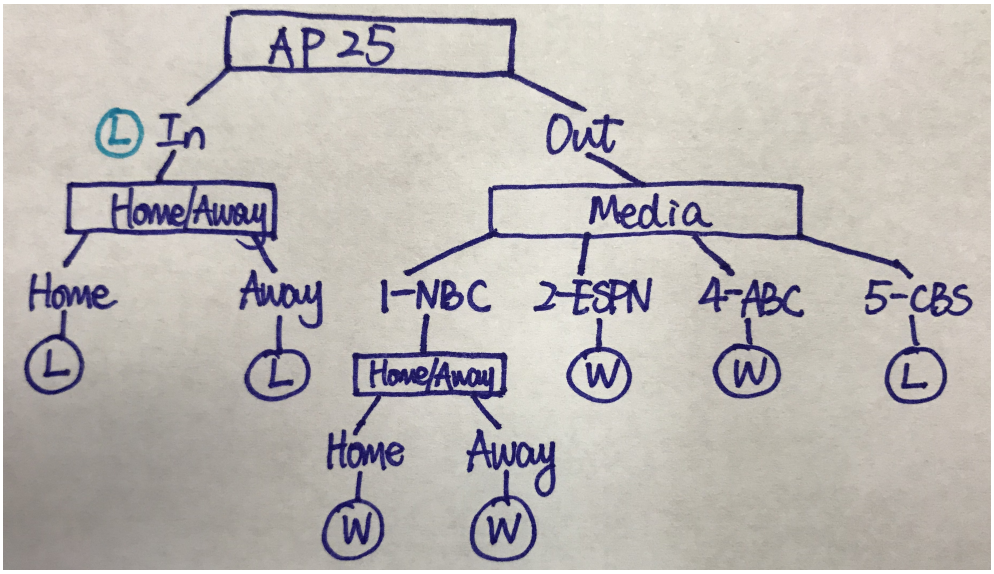

Figure 6: ID4.5-Final Decision Tree

3.2 Label testing sets based on the ID4.5 tree

We don't have training data given situation when opponent out AP25 at media 3-FOX, so here the given prediction based on out AP25.

ID	AP25	Media	Home/Away	Label	Truth
25	Out	1-NBC	Home	Win	Win
26	In	1-NBC	Home	Lose	Lose
27	Out	2-ESPN		Win	Win
28	Out			Win	Win
29	Out	1-NBC	Home	Win	Win
30	Out	4-ABC		Win	Win

3.3 Evaluation

Actual/Prediction	Win	Lose	Total
Win	5 (TP)	0 (FN)	5
Lose	0 (FP)	1 (TN)	1
Total	5	1	6

1. $Accuracy = \frac{TP+TN}{All} = \frac{6}{6} = 1.$
2. $Precision = \frac{TP}{TP+FP} = \frac{5}{5} = 1.$
3. $Recall = \frac{TP}{TP+FN} = \frac{5}{5} = 1.$
4. $F_1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*1*1}{1+1} = 1.$

3.4 Predictions

ID	AP25	Media	Home/Away	Prediction
31	In		Home	Lose
32	Out	1-NBC	Home	Win
33	Out	1-NBC	Home	Win
35	Out	1-NBC	Home	Win

4 Question 3 (20 points)

CART model, a decision tree model using "$\Delta gini$"

1. Construct a decision tree based on the training set (24 games).
2. Use the decision tree to predict labels of instances in the testing set (6 games) based on their attributes.
3. Calculate Accuracy, Precision, Recall, and F-1 score on the testing set, given the ground-truth labels of the 6 games.
4. Predict labels of the future 4 games using the decision tree.

Solutions:

4.1 Construct decision tree - CART

4.1.1 Step 1:

1. We want to predict whether a game is win or loss based on 3 attributes.
2. Y: (Win*14, Lose*10), $Gini(Y) = 1 - \sum_{i=1}^{k} p_i^2 = 1 - (\frac{14}{24})^2 - (\frac{10}{24})^2 = 0.4861.$
3. $X_{HomeAway}$: (Home*15, Away*9); Home: (Win*10, Lose*5), Away: (Win*4, Lose*5).

4. $Gini(Y|X_{HomeAway}) = \frac{15}{24}(1 - (\frac{10}{15})^2 - (\frac{5}{15})^2) + \frac{9}{24}(1 - (\frac{4}{9})^2 - (\frac{4}{9})^2) = 0.4630$.
5. $\Delta Gini(Y|X_{HomeAway}) = 0.4861 - 0.4630 = 0.0231$.
6. Similar as above, we can get $\Delta Gini(Y|X_{AP25}) = 0.0729$ and $\Delta Gini(Y|X_{Media}) = 0.1051$.
7. Then we choose the attributes with maximum $\Delta gini$ which is X_{Media}.
8. Then we will have our first tree in Figure 7.

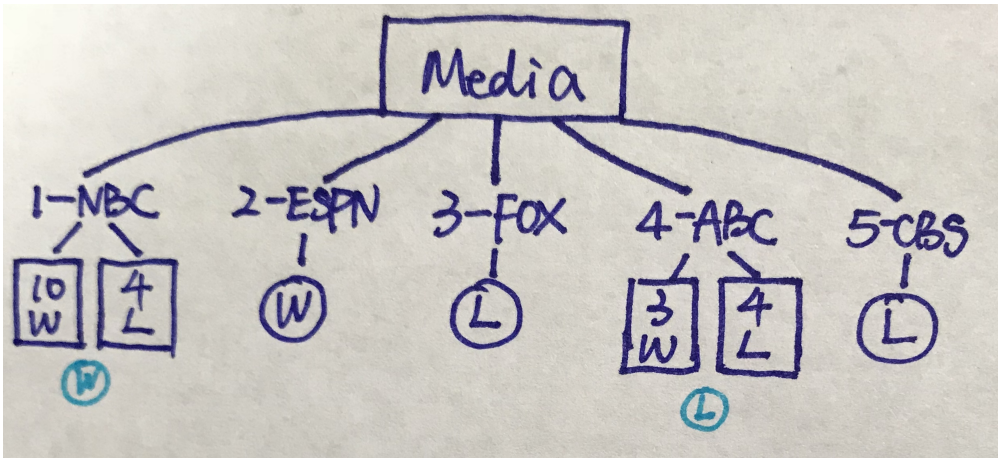

Figure 7: CART-Tree1

4.1.2 Step 2:

1. Since we already picked Media as our first attributes, we can tell from Tree1, we need add extra attribute into our decision tree.
2. Given media at 1-NBC, 14 events, Y: (Win*10, Lose*4), $Gini(Y|NBC) = 1 - (\frac{10}{14})^2 - (\frac{4}{14})^2 = 0.4082$
3. $X_{HomeAway}$: (Home*13, Away*1); Home: (Win*9, Lose*4), Away: (Win*1).
4. $Gini(Y|X_{HomeAway}, NBC) = \frac{13}{14}(1 - (\frac{9}{13})^2 - (\frac{4}{14})^2) + \frac{1}{14}(1 - (\frac{1}{1})^2) = 0.3956$.
5. $\Delta Gini(Y|X_{HomeAway}, NBC) = 0.4082 - 0.3956 = 0.0126$.
6. Similar as above, we can get $\Delta Gini(Y|X_{AP25}, NBC) = 0.0368$.
7. Given media at 4-ABC, we total have 7 events, Y: (Win*3, Lose*4)
8. $\Delta Gini(Y|X_{HomeAway}, ABC) = 0$ and $\Delta Gini(Y|X_{AP25}, ABC) = 0.1469$.
9. We choose attributes for maximum information gain, both X_{AP25} for 1-NBC and 4-ABC.

10. Then we will have our second tree in Figure 8.

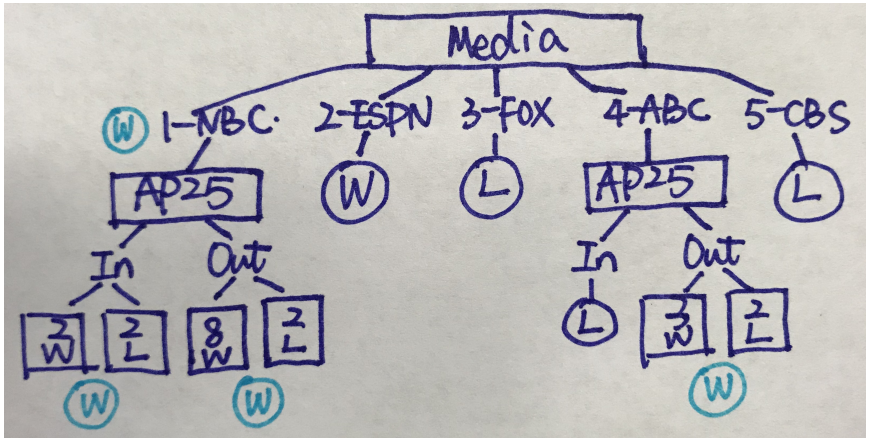

Figure 8: CART -Tree2

4.1.3 Step 3:

1. Given media at 1-NBC, opponent in AP25, 4 events, Y: (Win*2, Lose*2), $Gini(Y|NBC, In) = 0.5$.
2. $Gini(Y|X_{HomeAway}, NBC, In) = 0.5$, $\Delta Gini(Y|X_{HomeAway}, NBC, In) = 0$.
3. Given media at 1-NBC, opponent out AP25, 10 events, Y: (Win*8, Lose*2), $Gini(Y|NBC, Out) = 0.32$.
4. $Gini(Y|X_{HomeAway}, NBC, Out) = 0.3111$ and $\Delta Gini(Y|X_{HomeAway}, NBC, Out) = 0.0089$.
5. Given Media at 4-ABC, opponent out AP25, 5 events, Y: (Win*3, Lose*2).
6. $\Delta Gini(Y|X_{HomeAway}, ABC, Out) = 0$.
7. Then we will have our final decision tree in Figure 9.
8. The answer provided here for 50/50 is based on it's root node.

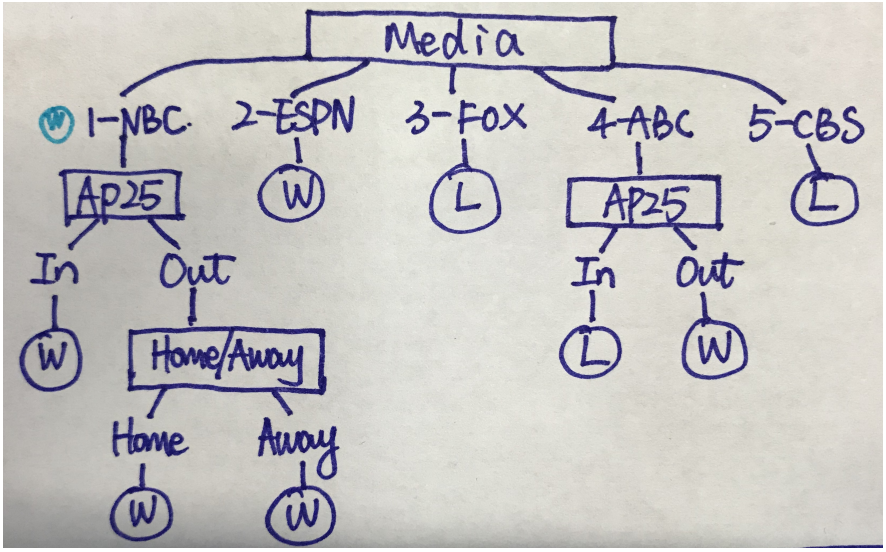

Figure 9: CART-Final Decision Tree

4.2 Label testing sets based on the CART

ID	Media	AP25	Home/Away	Label	Truth
25	1-NBC	Out	Home	Win	Win
26	1-NBC	In	Home	Win	Lose
27	2-ESPN			Win	Win
28	3-FOX			Lose	Win
29	1-NBC	Out	Home	Win	Win
30	4-ABC	Out	Away	Win	Win

4.3 Evaluation

Actual/Prediction	Win	Lose	Total
Win	4 (TP)	1 (FN)	5
Lose	1 (FP)	0 (TN)	1
Total	5	1	6

1. $Accuracy = \frac{TP+TN}{All} = \frac{4}{6} = 0.667.$
2. $Precision = \frac{TP}{TP+FP} = \frac{4}{5} = 0.8.$
3. $Recall = \frac{TP}{TP+FN} = \frac{4}{5} = 0.8.$
4. $F_1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*0.8*0.8}{0.8+0.8} = 0.8.$

4.4 Predictions

ID	Media	AP25	Home/Away	Prediction
31	1-NBC	In		Win
32	1-NBC	Out	Home	Win
33	1-NBC	Out	Home	Win
35	1-NBC	Out	Home	Win

5 Question 4 (30 points)

Naïve Bayes model.

- For each instance in the testing set (6 games), use NaiÌĺve Bayes to predict the label based on the training set (24 games).
- Calculate Accuracy, Precision, Recall, and F-1 score on the testing set, given the ground- truth labels of the 6 games.
- For each instance in the predicting set (4 games), use NaiÌĺve Bayes to predict the label based on the training set (24 games).

Solutions:

5.1 Label testing sets based on Naïve Bayes model

Prior probability: $p(win) = \frac{7}{12} = 0.5833$, $p(win) = \frac{5}{12} = 0.4166$.

5.1.1 Game 25

1. With opponent out AP25 at 1-NBC and Home, $P(Home|Win) = \frac{10}{14}$ and $P(Home|Lose) = \frac{5}{10}$
2. $P(Out|Win) = \frac{12}{14}$, $P(Out|Lose) = \frac{5}{10}$, $P(NBC|Win) = \frac{10}{14}$ and $P(NBC|Lose) = \frac{4}{10}$
3. Posterior probability: $P(X = Home, Out, NBC) = \frac{15}{24} * \frac{17}{24} * \frac{14}{24} = 0.2582$.
4. $P(X|Win) = \frac{10}{14} * \frac{12}{14} * \frac{10}{14} = 0.4373$ and $P(X|Lose) = \frac{5}{10} * \frac{5}{10} * \frac{4}{10} = 0.1$.
5. $P(Win|X) = P(X|Win) * P(Win)/P(X) = 0.4373 * 0.5833/0.2582 = 0.9880$.
6. $P(Lose|X) = P(X|Lose) * P(Lose)/P(X) = 0.1 * 0.4166/0.2582 = 0.1613$.
7. $P(Win|X) > P(Lose|X)$, so Game 25 Win.

5.1.2 Game 26

1. With opponent in AP25 at 1-NBC and Home, similar as before.
2. $P(Win|X) = 0.4$.
3. $P(Lose|X) = 0.3919$.
4. $P(Win|X) > P(Lose|X)$, so Game 26 Win.

5.1.3 Game 27

1. With opponent out AP25 at 2-ESPN and Away, similar as before.
2. $P(Win|X) = 0.9196$.
3. $P(Lose|X) = 0$.
4. $P(Win|X) > P(Lose|X)$, so Game 27 Win.

5.1.4 Game 28

1. With opponent out AP25 at 3-FOX and Away, similar as before
2. $P(Win|X) = 0$.
3. $P(Lose|X) = 0.9383$.
4. $P(Win|X) < P(Lose|X)$, so Game 28 Lose.

5.1.5 Game 29

1. Same as Game 25, so Game 29 Win.

5.1.6 Game 30

1. With opponent out AP25 at 4-ABC and Away, similar as before
2. $P(Win|X) = 0.3941$.
3. $P(Lose|X) = 0.5362$.
4. $P(Win|X) < P(Lose|X)$, so Game 30 Lose.

5.2 Evaluation

Actual/Prediction	Win	Lose	Total
Win	3 (TP)	2 (FN)	5
Lose	1 (FP)	0 (TN)	1
Total	4	2	6

1. $Accuracy = \frac{TP+TN}{All} = \frac{3}{6} = 0.5$.
2. $Precision = \frac{TP}{TP+FP} = \frac{3}{4} = 0.75$.
3. $Recall = \frac{TP}{TP+FN} = \frac{3}{5} = 0.6$.
4. $F_1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*0.75*0.6}{0.75+0.6} = 0.6667$.

5.3 Predictions

5.3.1 Game 31

1. With opponent in AP25 at 1-NBC and Home, same as game 26
2. Game 31 Win.

5.3.2 Game 32

1. With opponent out AP25 at 1-NBC and Home, same as game 25
2. Game 32 Win.

5.3.3 Game 33

1. With opponent out AP25 at 1-NBC and Home, same as game 25
2. Game 33 Win.

5.3.4 Game 35

1. With opponent out AP25 at 1-NBC and Home, same as game 25
2. Game 35 Win.

6 (5 points) Give your conclusion on which of the four models is the best.

C4.5 Model is the best model for this data set because it have the highest accuracy, precision, recall and F_1-score among the four models we used.