Diane Nguyen
Class Assignment 4
MSBA 645
4/4/19

Use the given RidingMower file to answer the following questions.

- **Split the initial dataset using a lot size of 19. Find the entropies of the root node and the entropies of the resulting child nodes. Find the information gain from the split. Compare the split decision using the lot size of 19 with that using an income of 59.7 (as computed on page 192). Which split decision would you take? Justify your answer. Show all steps in your computation.**

Household	Income	Lot_Size	Ownership
1	60	18.4	Owner
2	85.5	16.8	Owner
3	64.8	21.6	Owner
4	61.5	20.8	Owner
5	87	23.6	Owner
6	110.1	19.2	Owner
7	108	17.6	Owner
8	82.8	22.4	Owner
9	69	20	Owner
10	93	20.8	Owner
11	51	22	Owner
12	81	20	Owner
13	75	19.6	Nonowner
14	52.8	20.8	Nonowner
15	64.8	17.2	Nonowner
16	43.2	20.4	Nonowner
17	84	17.6	Nonowner
18	49.2	17.6	Nonowner
19	59.4	16	Nonowner
20	66	18.4	Nonowner
21	47.4	16.4	Nonowner
22	33	18.8	Nonowner
23	51	14	Nonowner
24	63	14.8	Nonowner

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i)$$

$$entropy\ (A) = -\sum_{k=1}^{m} p_k\ log_2(p_k)$$

- H(X) = entropy
- $P(x_i)$ = the proportion of observations in a rectangle X which belong to class I (out of n classes).
- Log_b = log with base 2
- The Entropy ranges between 0 (for most pure) and $log_2(n)$ (equal representation of classes)

Steps to solving entropy according to the book:

- Obtain overall impurity measure (weighted avg. of individual rectangles)
- At each successive stage, compare this measure across all possible splits in all variables
- Choose the split that reduces impurity the most
- Chosen split points become nodes on the tree

Steps to solving entropy formula in excel:

1. calculate the initial root node impurity by: a. get the values for each class of impurity; b. use the negative values * sum of the total impurity of the root node
2. Now you must calculate the impurity of the split in the data. a. get the value for each nodes impurity; b. weighted average of the impurity; c. get the value for change of entropy (done by subtracting the root node entropy be each of the weighted splits

Below is the entropy for the root nodes:

Log Bits: 1		Total # of options available: 2			**for each calculation:**
Impurity	**n**	**xi**	**xi * logb(xi)**		calulate entropy for i iteration
Class 1: Owner	12	0.5	-0.5		calulate entropy for i iteration
Class 2: Nonowner	12	0.5	-0.5		then put a negative and sum the entropies
TOTAL	24		1	Has a lot of impurity	

X_i = n/TOTAL → 12/24=0.5

Now we must Split using a Lot Size of 19

Household	Income	Lot_Size	Ownership	Split
3	64.8	21.6	Owner	split1
4	61.5	20.8	Owner	split1
5	87	23.6	Owner	split1
6	110.1	19.2	Owner	split1
8	82.8	22.4	Owner	split1
9	69	20	Owner	split1
10	93	20.8	Owner	split1
11	51	22	Owner	split1
12	81	20	Owner	split1
13	75	19.6	Nonowner	split1
14	52.8	20.8	Nonowner	split1
16	43.2	20.4	Nonowner	split1
1	60	18.4	Owner	split2
2	85.5	16.8	Owner	split2
7	108	17.6	Owner	split2
15	64.8	17.2	Nonowner	split2
17	84	17.6	Nonowner	split2
18	49.2	17.6	Nonowner	split2
19	59.4	16	Nonowner	split2
20	66	18.4	Nonowner	split2
21	47.4	16.4	Nonowner	split2
22	33	18.8	Nonowner	split2
23	51	14	Nonowner	split2
24	63	14.8	Nonowner	split2

Below is the entropy on the split on Lot Size of 19

Impurity	n	xi	xi * logb(xi)	
Class 1: Owner	9	0.75	-0.31128	
Class 2: Nonowner	3	0.25	-0.5	
TOTAL	12		0.811278	still impure

Impurity	n	xi	xi * logb(xi)	
Class 1: Owner	3	0.25	-0.5	
Class 2: Nonowner	9	0.75	-0.31128	
TOTAL	12		0.811278	still impure
change in entropy:	-0.18872	The amount of entropy has been reduced		

X_i above = n/TOTAL => 9/12=0.75; 3/12=0.25

information gain is entropy of the root - weighted sum of child entropies
1-(((12/24)*0.81)+((12/24)*0.81))
0.190

Now we are splitting on Income of 59.7

(here split by income of 59.7, then sort the data based on Household)

Household	Income	Lot_Size	Ownership	Split
1	60	18.4	Owner	split1
2	85.5	16.8	Owner	split1
3	64.8	21.6	Owner	split1
4	61.5	20.8	Owner	split1
5	87	23.6	Owner	split1
6	110.1	19.2	Owner	split1
7	108	17.6	Owner	split1
8	82.8	22.4	Owner	split1
9	69	20	Owner	split1
10	93	20.8	Owner	split1
12	81	20	Owner	split1
13	75	19.6	Nonowner	split1
15	64.8	17.2	Nonowner	split1
17	84	17.6	Nonowner	split1
20	66	18.4	Nonowner	split1
24	63	14.8	Nonowner	split1
11	51	22	Owner	split2
14	52.8	20.8	Nonowner	split2
16	43.2	20.4	Nonowner	split2
18	49.2	17.6	Nonowner	split2
19	59.4	16	Nonowner	split2
21	47.4	16.4	Nonowner	split2
22	33	18.8	Nonowner	split2
23	51	14	Nonowner	split2

Impurity	n	xi	xi * logb(xi)	
Class 1: Owner	11	0.6875	-0.37164	
Class 2: Nonowner	5	0.3125	-0.5244	
TOTAL	16		0.896038	still impure

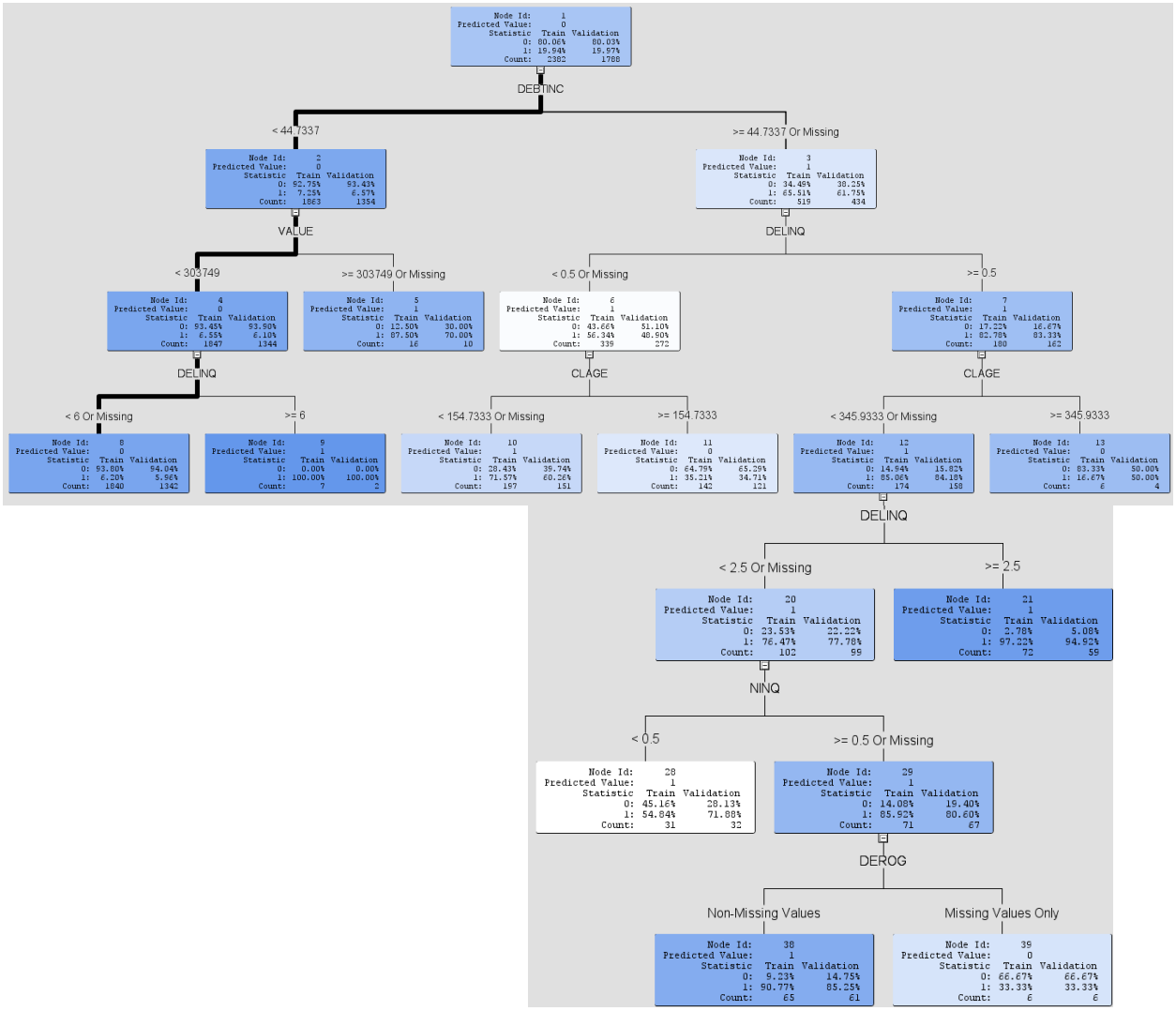
Impurity	n	xi	xi * logb(xi)	
Class 1: Owner	1	0.125	-0.375	
Class 2: Nonowner	7	0.875	-0.16856	
TOTAL	8		0.543564	still impure

information gain = entropy of the root - weighted sum of child entropies
1-(((16/24)*0.89)+((8/24)*0.54))
0.227

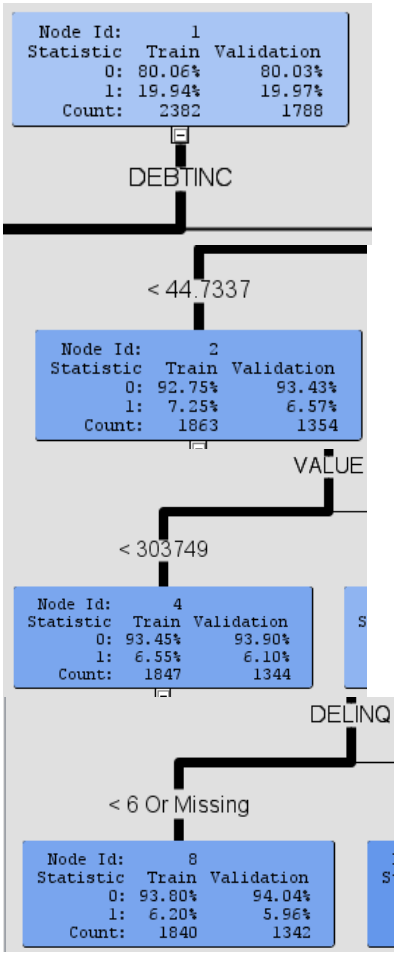
The split based on income <59.7 which equals 0.543564 because it gives you the lowest entropy value which means that it is most pure.

- **Use the decision tree built on the home equity loan to determine how the following two cases would be classified. List the nodes traversed in order of visitation. If you decide to list the individual nodes, make sure you include the split information if it is not a leaf node.**

Obs #	BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
5364	0	30500	124225	181129	DebtCon	Mgr	11	0	0	189.4392	2	28	43.16839
5365	1	30500	18375	61738	DebtCon	Mgr	20	3	5	141.7031	0	28	45.35479

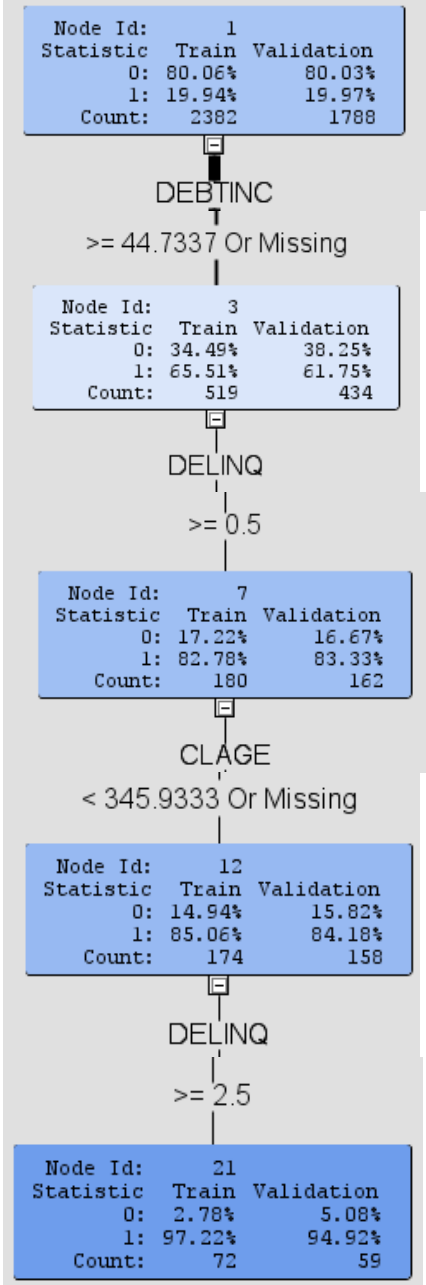


For obs # 5364, the decision tree you would start with the DEBTINC, you would follow it down the left hand brand because the DEBTINC was <44.7337, you then would follow VALUE for the branch <303749 and finally follow the < 6 Or Missing for DELINQ

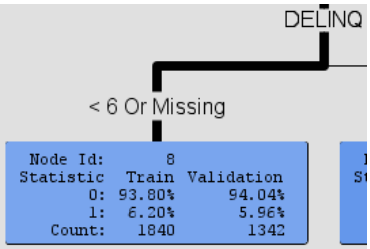


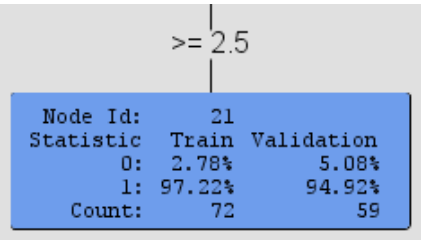
For obs # 5365, the decision tree you would start with the DEBTINC >44.7337, next you would go to the DELINQ for >=0.5, then follow the CLAGE branch for <345.9333, then you would follow DELINQ for >=2.5

Obs #	BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
5364	0	30500	124225	181129	DebtCon	Mgr	11	0	0	189.4392	2	28	43.16839
5365	1	30500	18375	61738	DebtCon	Mgr	20	3	5	141.7031	0	28	45.35479



The Obs # 5364 would be classified as a zero according to the decision tree and OBS # 5365 would be classified as a 1.





- **Identify the same cases with rules. List the rule used in each case.**

Obs #	BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
5364	0	30500	124225	181129	DebtCon	Mgr	11	0	0	189.4392	2	28	43.16839
5365	1	30500	18375	61738	DebtCon	Mgr	20	3	5	141.7031	0	28	45.35479

Go to view→model→ node rules
For Obs # 5364

```
*-------------------------------------------------------------*
 Node = 8
*-------------------------------------------------------------*
if VALUE < 303749
AND DELINQ < 6 or MISSING
AND DEBTINC < 44.7337
then
 Tree Node Identifier   = 8
 Number of Observations = 1840
 Predicted: BAD=1 = 0.06
 Predicted: BAD=0 = 0.94
```

For Obs # 5365

```
*-------------------------------------------------------------*
 Node = 21
*-------------------------------------------------------------*
if DELINQ >= 2.5
AND DEBTINC >= 44.7337 or MISSING
AND CLAGE < 345.933 or MISSING
then
 Tree Node Identifier   = 21
 Number of Observations = 72
 Predicted: BAD=1 = 0.97
 Predicted: BAD=0 = 0.03

*-------------------------------------------------------------*
```