Assignment 2

Chen Siyi A0194556R Gao Haochun A0194525Y He Yuan A0211297H
Wang Pei A0194486M Wang Zi A0194504E

10/18/2020

Question 1

1.a

In the whole dataset, the proprtions of the samples of yes and no are:

$$p_y = \frac{6}{10} = \frac{3}{5} \quad and \quad p_n = \frac{4}{10} = \frac{2}{5}$$

$$Gini = 1 - (p_y \times p_y + p_n \times p_n) = 0.48$$

1.b

Firstly, we need to sort the dataset according to the attribute *Year Born*

Year Born	Bonus above median
63	YES
65	YES
70	YES
72	YES
75	NO
82	NO
83	NO
84	NO
86	YES
88	YES

There are two possible spliting cases:

Case 1: *Year Born* <= 72 and *Year Born* > 72

Left node: number of samples = 4, including 4 samples of "yes" and 0 samples of "no"

$$p_y = \frac{4}{4} = 1 \quad and \quad p_n = \frac{0}{4} = 0$$

Hence, the Gini index of the left node is:

$$Gini(left) = -1 - (p_y \times p_y + p_n \times p_n) = 0$$

Right node: number of samples = 6, including 2 samples of "yes" and 4 samples of "no"

$$p_y = \frac{2}{6} = \frac{1}{3} p_n = \frac{4}{6} = \frac{2}{3}$$

Hence, the Gini index of the left node is:

$$Gini(right) = -1 - (p_y \times p_y + p_n \times p_n) = \frac{4}{9}$$

The total avarage Gini index and the reduction in terms of Gini index:

$$Gini(left, right) = \frac{4}{10} \times Gini(left) + \frac{6}{10} \times Gini(right) = 0.2667$$

$$The reduction in terms of Gini index = 0.48 - Gini(left, right) = 0.2133$$

Case 2: Year Born<=84 and Year Born>84

Left node: number of samples = 8, including 4 samples of "yes" and 4 samples of "no"

$$p_y = \frac{4}{8} = \frac{1}{2} p_n = \frac{4}{8} = \frac{1}{2}$$

Hence, the Gini index of the left node is:

$$Gini(left) = -1 - (p_y \times p_y + p_n \times p_n) = \frac{1}{2}$$

Right node: number of samples = 2, including 2 samples of "yes" and 0 samples of "no"

$$p_y = \frac{2}{2} = 1 p_n = \frac{0}{2} = 0$$

Hence, the Gini index of the left node is:

$$Gini(right) = -1 - (p_y \times p_y + p_n \times p_n) = 0$$

The total avarage Gini index and the reduction in terms of Gini index:

$$Gini(left, right) = \frac{8}{10} \times Gini(left) + \frac{2}{10} \times Gini(right) = 0.4$$

$$The reduction in terms of Gini index = 0.48 - Gini(left, right) = 0.08$$

Therfore, the maximum reduction is 0.08, which is the spit at Year Born<=84 and Year Born>84.

1.c

Since *Job Type* is a categorical discrete attribute with 5 values, it has

$$2^{(5-1)} - 1 = 15$$

possible splits:

Case	Split
1	1 versus 2,3,4,5
2	2 versus 1,3,4,5
3	3 versus 1,2,4,5
4	4 versus 1,2,3,5
5	5 versus 1,2,3,4
6	1,2 versus 3,4,5
7	1,3 versus 2,4,5
8	1,4 versus 2,3,5
9	1,5 versus 2,3,4
10	2,3 versus 1,4,5
11	2,4 versus 1,3,5
12	2,5 versus 1,3,4
13	3,4 versus 1,2,5
14	3,5 versus 1,2,4
15	4,5 versus 1,2,3

1.d

Since *Education Level* is an ordinal discrete attribute with 4 values, it has

$$4 - 1 = 3$$

possible splits:

Case	Split
1	1 versus 2,3,4
2	1,2 versus 3,4
3	1,2,3 versus 4

1.e

First method:

Calculate the mean of *Year Born* of other 10 employees, and fill the *Year Born* of No.11 Employee with this mean value.

Second method:

Use all the attributes (except *Year Born*) of the original 10 employees to build a regression model with respect to *Year Born*. Then, we can use this model to predict the *Year Born* of No.11 employee based on its other attributes.

Question 2

2.a

In the whole dataset, the proprtions of the samples of yes and no are:

$$p_y = \frac{6}{12} = \frac{1}{2} \quad and \quad p_n = \frac{6}{12} = \frac{1}{2}$$

Hence, the entropy of the training samples is:

$$Entropy(root) = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n) = 1$$

2.b

Firstly, we need to sort the dataset according to the attribute "Income"

INCOME	GOOD-CREDIT
21	YES
22	NO
29	NO
34	NO
45	NO
49	NO
72	YES
81	NO
100	YES
105	YES
125	YES
180	YES

There are four possible spliting cases:

Case 1: income<=81 and income>81

Left node: number of samples = 8, including 2 samples of "yes" and 6 samples of "no"

$$p_y = \frac{2}{8} = \frac{1}{4} p_n = \frac{6}{8} = \frac{3}{4}$$

Hence, the entropy of the left node is:

$$Entropy(left) = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n)$$

Right node: number of samples = 4, including 4 samples of "yes" and 0 samples of "no"

$$p_y = 1 p_n = 0$$

Hence, the entropy of the right node is:

$$Entropy(right) = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n)$$

The total avarage Entropy and the information gain:

$$I_G(left, right) = \frac{8}{12} \times Entropy(left) + \frac{4}{12} \times Entropy(right) = 0.541$$

$$Information\ Gain = 1 - I_G(left, right) = 0.459$$

Case 2: income<=72 and income>72

Left node: number of samples = 7, including 2 samples of "yes" and 5 samples of "no"

$$p_y = \frac{2}{7} p_n = \frac{5}{7}$$

Hence, the entropy of the left node is:

$$Entropy(left) = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n)$$

Right node: number of samples = 5, including 4 samples of "yes" and 1 samples of "no"

$$p_y = \frac{4}{5} p_n = \frac{1}{5}$$

Hence, the entropy of the right node is:

$$Entropy(right) = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n)$$

The total avarage Entropy and the information gain:

$$I_G(left, right) = \frac{7}{12} \times Entropy(left) + \frac{5}{12} \times Entropy(right) = 0.804$$

$$Information\ Gain = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n) = 0.196$$

Case 3: income<=21 and income>21

Left node: number of samples = 1, including 1 samples of "yes" and 0 samples of "no"

$$p_y = 1 p_n = 0$$

Hence, the entropy of the left node is:

$$Entropy(left) = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n) = 0$$

Right node: number of samples = 11, including 5 samples of "yes" and 6 samples of "no"

$$p_y = \frac{5}{11} p_n = \frac{6}{11}$$

Hence, the entropy of the right node is:

$$Entropy(right) = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n)$$

The total avarage Entropy and the information gain:

$$I_G(left, right) = \frac{1}{12} \times Entropy(left) + \frac{11}{12} \times Entropy(right) = 0.9119$$

$$Information\ Gain = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n) = 0.09$$

Case 4: income<=49 and income>49

Left node: number of samples = 6, including 5 samples of "yes" and 1 samples of "no"

$$p_y = \frac{5}{6} p_n = \frac{1}{6}$$

Hence, the entropy of the left node is:

$$Entropy(left) = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n)$$

Right node: number of samples = 6, including 1 samples of "yes" and 5 samples of "no"

$$p_y = \frac{1}{6} p_n = \frac{5}{6}$$

Hence, the entropy of the right node is:

$$Entropy(right) = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n)$$

The total avarage Entropy and the information gain:

$$I_G(left, right) = \frac{7}{12} \times Entropy(left) + \frac{5}{12} \times Entropy(right) = 0.650$$

$$Information\ Gain = -(p_y \times \log_2 p_y + p_n \times \log_2 p_n) = 0.35$$

The information gain = 0.459 in case1 is the largest, so the best cut-off point for splitting is the income between 81 and 100.

2.c

Let's select the attribute "Income" as the first attribute. As discussed in 2.b, we know that we should split the dataset into "less or equal to 81" and "more than 81".

Since the entropy(right) is already 0, we only need to consider how to continute split the left side of the decision tree.

We try to select the Age as the second attribute of the decision tree, and then sort the samples in the left node by the attribute "AGE"

AGE	GOOD-CREDIT
25	NO
29	NO
35	NO
37	NO
45	NO
50	NO
53	YES
58	YES

It is clear that we should split samples into "age is less or equal to 50" and "age is more than 50" (the number 50 can be repalced by any number in [50,53)), since both the entropy in the left and right will be 0 and the information gain would maximize.

Question 3

3.a

No market survey: (Unit in $\$1x10^4$)

	FM(1/3)	UM(2/3)	Expectation
Major renovation	250	-80	30

	FM(1/3)	UM(2/3)	Expectation
Minor renovation	100	-20	20
Do nothing	0	0	0

Regret table:

	FM(1/3)	UM(2/3)	Oppo.loss
Major	0	80	160/3
Minor	150	20	190/3
Do nothing	250	0	250/3

Hence, **Major renovation** can minimize the expected opportunity loss to \$ $\frac{160}{3} \times 10^4 = 5.33333 \times 10^5$.

3.b

$$EVWOI = 250 \times \frac{1}{3} - 80 \times \frac{2}{3} = 30$$

3.c

$$EVWPI = 250 \times \frac{1}{3} + 0 \times \frac{2}{3} = \frac{250}{3}$$

$$EVOPI = EVWPI - EVWOI = \frac{160}{3} \times 10^4 = 5.33333 \times 10^5$$

3.d

Let

$$P(FM) = \alpha \ \ and \ \ P(UM) = 2P(FM) = 2\alpha$$

Given:

$$P(P|FM) = 0.8, P(P|UM) = 0.4, P(N|FM) = 0.2, P(N|UM) = 0.6$$

Total probability:

$$P(P) = P(FM)P(P|FM) + P(UM)P(P|UM) = 1.6\alpha$$
$$P(N) = P(FM)P(N|FM) + P(UM)P(N|UM) = 1.4\alpha$$

Combined with $P(P) + P(N) = 1$:

$$P(P) = \frac{8}{15} \ \ and \ \ P(N) = \frac{7}{15}$$

Probability conditioned on Positive result:

$$P(FM|P) = \frac{P(FM)P(P|FM)}{P(P)} = \frac{3}{2}\alpha$$
$$P(UM|P) = \frac{P(UM)P(P|UM)}{P(P)} = \frac{3}{2}\alpha$$

Combined with $P(FM|P) + P(UM|P) = 1$:

$$P(UM|P) = P(FM|P) = \frac{1}{2}$$

Probability conditioned on Negative result:

$$P(FM|N) = \frac{P(FM)P(N|FM)}{P(N)} = \frac{3}{7}\alpha$$
$$P(UM|N) = \frac{P(UM)P(N|UM)}{P(N)} = \frac{18}{7}\alpha$$

Combined with $P(FM|N) + P(UM|N) = 1$:

$$P(UM|N) = \frac{6}{7} \ \ and \ \ P(FM|N) = \frac{1}{7}$$

Free market survey with Positive result(8/15):

	FM(1/2)	UM(1/2)	Expectation
Major renovation	250	-80	85
Minor renovation	100	-20	40
Do nothing	0	0	0

Free market survey with Negative result(7/15):

	FM(1/7)	UM(6/7)	Expectation
Major renovation	250	-80	-230/7
Minor renovation	100	-20	-20/7
Do nothing	0	0	0

$$EVWSI = 85 \times \frac{8}{15} + 0 \times \frac{7}{15} = \frac{136}{3}$$
$$EVOSI = EVWSI - EVWOI = \frac{46}{3} \times 10^4 = 1.53333 \times 10^5$$

Conclusion

As EVOSI ($\frac{46}{3} \times 10^4$) > Cost of survey (5×10^4), the company should do market survey.

To maximize return, after doing market research, if the survey result is positive, it should do Major renovation, if the result is negative, it should do nothing.

$$Expected \ \ return = \frac{136}{3} - 5 = \frac{121}{3} \times 10^4 = 4.03333 \times 10^5$$