Decision Tree Analysis

Author: Han Jiatong

Discussion Question 1:

Employee No	Education Level	Job Type	Year Born (19xx)	Gender	Years Prior	Bonus above median
1	3	1	84	Male	1	no
2	1	5	72	Female	15	yes
3	1	1	75	Female	12	no
4	2	2	70	Female	17	yes
5	3	1	82	Male	5	no
6	3	1	86	Female	0	yes
7	3	1	83	Female	3	no
8	3	1	88	Male	2	yes
9	1	3	65	Male	24	yes
10	4	4	63	Male	32	yes

The variables in the dataset are:

- *Education Level:* a categorical variable with categories 1 (A level), 2 (polytechnic), 3 (bachelor's degree), 4 (post-graduate degree).
- *Job Type:* a categorical variable indicating the type of job the employee holds: 1 (management), 2 (sales), 3 (administration), 4 (office), 5 (miscellaneous support)
- *Year Born:* year employee was born, a continuous variable.
- *Gender:* Male or Female.
- *Years Prior:* number of years of work experience at another bank prior to working at ABC Bank, a continuous variable.
- *Bonus above median:* class label with value is "yes" if the employee receives a year-end bonus that is above the median bonus of the previous year, "no" otherwise.

(a) How heterogeneous are the samples in the dataset? Compute using the Gini index.

Gini index is defined to be $Gini(q) = 1 - \sum_h P_h^2$

From the dataset,

$$P_{above} = \frac{6}{10} = \frac{3}{5}$$

$$P_{below} = 1 - P_{above} = \frac{2}{5}$$

So the Gini index is calculated as

$$1 - (P_{above}^2 + P_{below}^2) = 1 - (\frac{3^2}{5^2} + \frac{2^2}{5^2}) = \frac{12}{25} = 0.48$$

(b) Using the values of the variable Year Born, what is the maximum reduction in the diversity that can be achieved? Measure the reduction in terms of Gini index.

After examining the data samples, we consider spliting where *Year-Born greater than or equal to 75* or *Year-Born less than 75*: (other splits are also possible but considered less rational)

Impurity after split:

$$I_G(q_1, q_2) = \frac{4}{10} \cdot Gini(q_1) + \frac{6}{10} \cdot Gini(q_2) = \frac{2}{5} \times 0 + \frac{3}{5} \times \frac{4}{9} = 0.267$$

So the achieved reduction measured by Gini index would be $Gini(q) - I_G(q_1, q_2) = 0.213$

(c) If we consider splitting the data according to the values of the variable Job Type, how many possible splits are there?

As *Job-Type* is a categorical variable, with $N = 5$ variations in total, the number of possible splits are $2^{N-1} - 1 = 2^4 - 1 = 15$.

(d) If we consider splitting the data according to the values of the variable Education Level, how many possible splits are there? List all these possible splits.

As *Education* is a categorical variable, taking values in $\{1, 2, 3, 4\}$, the possible splits are listed as follows:

$$\begin{aligned}
1 \quad & versus \quad 2, 3, 4 \\
2 \quad & versus \quad 1, 3, 4 \\
3 \quad & versus \quad 1, 2, 4 \\
4 \quad & versus \quad 1, 2, 3 \\
1, 2 \quad & versus \quad 3, 4 \\
1, 3 \quad & versus \quad 2, 4 \\
1, 4 \quad & versus \quad 2, 3
\end{aligned}$$

In total, there are 7 possible splits.

(e) Suppose the information from Employee No. 11 is now available, but with its value for the variable Year Born missing. Suggest 2 possible ways to handle the missing value so that this new data sample can be included for building the classification tree.

Approach 1: Use the **Identification** method

As the *Year-born* attribute is continuous, to encode and identify the missing data, we assign a value of -1 to the new data sample.

Approach 2: Use the **Substitution** method

We can substitute the missing value with the mean of the other observations, which is calculated as:

$$Mean = \frac{\sum_{i=1}^{10} Yearborn[i]}{10} \approx 77$$

Thus the value 77 is assigned to the new data sample as the year born.

Discussion Question 2:

AGE	PROF-EXP	INCOME	UNIVERSITY	GOOD-CREDIT
25	1	49	0	NO
29	5	45	1	NO
34	9	180	1	YES
35	8	125	0	YES
35	9	100	1	YES
35	10	81	0	NO
37	13	29	1	NO
45	19	34	1	NO
50	24	22	1	NO
53	27	72	1	YES
58	15	21	1	YES
65	39	105	0	YES

AGE, years of professional experience (PROF-EXP), annual INCOME (thousands of $) are continuous attributes, while UNIVERSITY = 0 if the client does not have a university degree, 1 otherwise.

A binary decision tree to differentiate between two groups of clients (GOOD-CREDIT YES vs GOOD-CREDIT NO) is proposed.

(a) What is the entropy of the training samples with respect to the classification?

Entropy index is defined to be $Entropy(q) = -\sum_h P_h \cdot log_2 P_h$

From the dataset,

$$P_{yes} = \frac{6}{12} = \frac{1}{2}$$

$$P_{no} = 1 - P_{yes} = \frac{1}{2}$$

The Entropy index is calculated as

$$-\sum_h P_h \cdot log_2 P_h = -(\frac{1}{2} log_2 \frac{1}{2} + \frac{1}{2} log_2 \frac{1}{2}) = -(log_2 \frac{1}{2}) = 1$$

So the Entropy index of the training samples is 1.

(b) Suppose INCOME is to be the first attribute used for splitting the root node of the decision tree. What is the best cut-off point for splitting that will maximize the information gain?

We shortlisted two possible cases of splits which are not obviously comparable in performance to determine which yields more information gain:

- Case 1: Split at income = 81 (thousand dollars)

  Node q_1 is the group of samples with income less than or equal to 81 (thousand dollars). Two samples (out of 8) are in Good-Credit. Its entropy index is computed as:

$$Entropy(q_1) = -\sum_h P_h \cdot log_2 P_h = -(\frac{2}{8} log_2 \frac{2}{8} + \frac{6}{8} log_2 \frac{6}{8}) = 0.81$$

Node q_2 is the group of samples with income greater than 81 (thousand dollars). We note that all the samples in q_2 are not in Good-Credit. Therefore its entropy is 0.

So the Impurity after split would be:

$$I_E(q_1, q_2) = \frac{8}{12} \cdot Entropy(q_1) + \frac{4}{12} \cdot Entropy(q_2) = 0 + \frac{2}{3} \times 0.81 = 0.54$$

$\therefore$ Information Gain = 1 - 0.54 = 0.46

- Case 2: Split at income = 72 (thousand dollars)

  Node q_1 is defined to be the group of samples with income less than 72 (thousand dollars). All sample points are **not in Good-Credit**, thus the entropy value is 0.

  Node q_2 has its sample points with income greater than or equal to 72 (thousand dollars). Only one sample (out of 7) is not in Good-credit. Thus its entropy index is calculated by:

$$Entropy(q_1) = -\sum_h P_h \cdot log_2 P_h = -(\frac{1}{7} \times log_2 \frac{1}{7} + \frac{6}{7} \times log_2 \frac{6}{7}) = 0.59$$

  So the Impurity after split would be:

$$I_E(q_1, q_2) = \frac{5}{12} \cdot Entropy(q_1) + \frac{7}{12} \cdot Entropy(q_2) = 0 + \frac{7}{12} \times 0.59 = 0.34$$

  $\therefore$ Information Gain = 1 - 0.34 = 0.66

We can see that 0.46 (case 1) $<$ 0.66 (case 2)

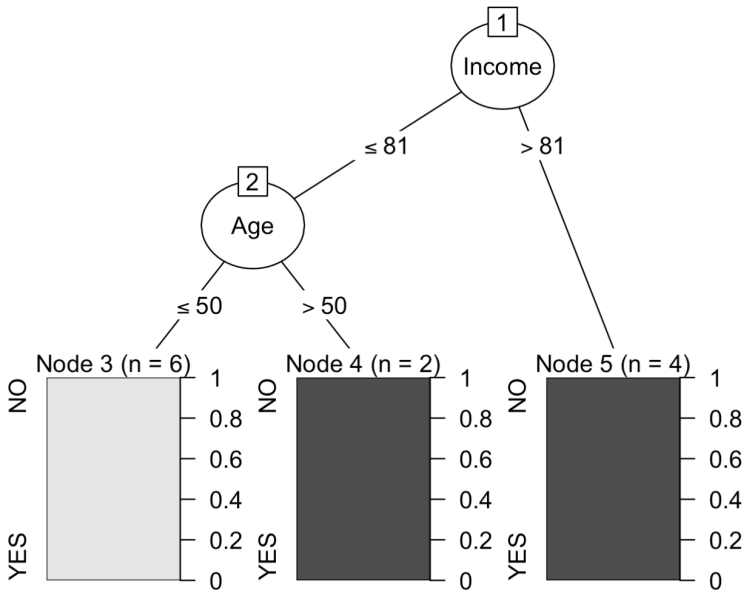
To conclude, since splitting at income = 72 gives more information gain, we decided to fix the final split point at income greater than or equal to 77 versus less than 77, which is the middle point between 72 and 81, to realize a more flexible and generalized model.

(c) Build a complete decision tree that classifies all the training data samples correctly.

We constructed a decision tree with the help of C5.0 package in R. The decision rule was:

1. if the *Income* is greater than 81 (thousand dollars), predict *Yes*;
2. if the *Income* is less than or equal to 81 (thousand dollars):
   - if the *Age* is less than or equal to 60, predict *No*;
   - if the *Age* is greater than 60, predict *Yes*;

The full decision tree is depicted as follows:



Obviously seen from the plot that the model classifies all the training samples correctly.

Discussion Question 3

Gold Food is considering the introduction of a new line of desserts. In order to produce the new line, the company is considering either a major renovation or a minor renovation of its current production facility. The following table shows the expected returns of the various alternatives that it can choose:

Alternative	Market condition	
	Favourable mkt	Unfavourable mkt
Major renovation	$2,500,000	-$800,000
Minor renovation	$1,000,000	-$200,000
Do nothing	$0	$0

Before making the final decision, Gold Food can conduct a marketing research survey at a cost of $50,000. The effectiveness of similar research surveys in the past in predicting the actual nature of the market is shown in the table below.

	Actual State of Nature	
Survey results	Favourable market (FM)	Unfavourable market (UM)
Positive (PS)	0.80	0.40
Negative (NS)	0.20	0.60

(for example, when the market condition is favorable, the probability of a positive outcome in the survey is 80%)

Assume that without any survey information, the probability of an unfavourable market is twice the probability of a favourable market.

(a) What is the decision that minimizes the expected opportunity loss (regret)?

Regret (opportunity cost) matrix:(in ten thousand dollars)

	Favorable.market	Unfavorable.market
Major renovation	0.000	80.000
Minor renovation	150.000	20.000
Do nothing	250.000	0.000
Probability	0.333	0.667

If *Major Renovation* is exercised, expected opportunity loss = (1/3) * 0 + (2/3) * 80 = 53.333;
Else if *Minor Renovation* is exercised, expected opportunity loss = (1/3) * 150 + (2/3) * 20 = 63.333;
Else *Nothing* is done, then the expected opportunity loss = (1/3) * 250 + (2/3) * 0 = 83.333;

$\therefore$ Choose *Major Renovation* to minimise expected opportunity loss.

(b) Determine the Expected Value with Original Information.

The expected values grid with original information is depicted as:

	Favorable.market	Unfavorable.market	Expected.Profits
Major renovation	250	-80	0.333(250) + 0.667(-80) = 30
Minor renovation	100	-20	0.333(100) + 0.667(-20) = 20
Do nothing	0	0	0.333(0) + 0.667(0) = 0

So the expected values (in ten thousand of dollars) for each course of action will be:

	Expected.Profits
Major renovation	30
Minor renovation	20
Do nothing	0

(c) Determine the Expected Value of Perfect Information.

	Favorable.market	Unfavorable.market
Major renovation	250.000	-80.000
Minor renovation	100.000	-20.000
Do nothing	0.000	0.000
Prob. of Favorability	0.333	0.667

(All the following metrics are in ten thousand of dollars)
Expected Value With Perfect Information (EVWPI) = (1/3) * 250 + (2/3) * 0 = 83.333

Expected Profit With Original Information (EVWOI) to maximize EV = 30

Thus the Expected Value of Perfect Information (EVPI) = EVWPI - EVWOI = 83.333 - 30 = 53.333

(d) Determine the Expected Value of Sample Information and the best course of action.

We have the **prior probability**:

$$P(FM) = \frac{1}{3} \quad P(UM) = \frac{2}{3}$$
$$P(PS|FM) = 0.80 \quad P(NS|FM) = 0.20$$
$$P(PS|UM) = 0.40 \quad P(NS|UM) = 0.60$$

Joint probability:

$$P(PS \cap FM) = P(FM) \cdot P(PS|FM) = \frac{4}{15} \approx 0.267$$
$$P(NS \cap FM) = P(FM) \cdot P(NS|FM) = \frac{1}{15} \approx 0.067$$
$$P(PS \cap UM) = P(UM) \cdot P(PS|UM) = \frac{4}{15} \approx 0.267$$
$$P(NS \cap UM) = P(UM) \cdot P(NS|UM) = 0.400$$

Marginal probabilities:

$$P(PS) = P(PS \cap FM) + P(PS \cap UM) = \frac{8}{15} \approx 0.534$$
$$P(NS) = P(NS \cap FM) + P(NS \cap UM) = \frac{7}{15} \approx 0.467$$

Posterior (revised) probabilities:

$$P(FM|PS) = P(FM \cap PS)/P(PS) = \frac{\frac{4}{15}}{\frac{8}{15}} = 0.5$$
$$P(UM|PS) = P(UM \cap PS)/P(PS) = \frac{\frac{4}{15}}{\frac{8}{15}} = 0.5$$
$$P(FM|NS) = P(FM \cap NS)/P(NS) = \frac{\frac{1}{15}}{\frac{7}{15}} = \frac{1}{7} \approx 0.143$$
$$P(UM|NS) = P(UM \cap NS)/P(NS) = \frac{0.4}{\frac{7}{15}} = \frac{6}{7} \approx 0.857$$

If we decide to conduct the market research:
(cost is 5 in ten thousand dollars)

- If the research outcome is *positive*:
  - Major renovation:
    E1 = (0.5)(250) + (0.5)(-80) = 85
  - Minor renovation:
    E2 = (0.5)(100) + (0.5)(-20) = 40
  - Do nothing:
    E3 = (0.5)(0) + (0.5)(0) = 0

  It is better to adopt **major renovation**.
- If the research outcome is *negative*:

- Major renovation:
  E4 = (1/7)(250) + (6/7)(-80) = -32.86
- Minor renovation:
  E5 = (1/7)(100) + (6/7)(-20) = -2.857
- Do nothing:
  E6 = (1/7)(0) + (6/7)(0) = 0

It is better to **do nothing**.

Therefore, the expected return would be:

$$E_{research} = P(PS) \cdot E1 + P(NS) \cdot E6 = \frac{8}{15} \times 85 + \frac{7}{15} \times 0 = 45.33$$

which is **40.33** if deducted by the cost.

Compared to: (no research conducted thus no information acquired)

$$E_{major} = 30 \quad E_{minor} = 20 \quad E_{nothing} = 0$$

Apparently the best course of action is to carry out the market research.