Tut 7: Decision Tree Models

June 2024

Classification Tree

1. Use **gain ratio** to determine which split is better:

   Split 1: Leaf $A = [20+, 15-]$; Leaf $B = [5+, 20-]$

   Split 2: Leaf $A = [10+, 2-]$; Leaf $B = [15+, 33-]$

   Remark: The larger "information gain" and "gain ratio", the better.

   Solution. Total, $Tbl(S) = [25+, 35-]$ implies $H(S) = -(\frac{25}{60}\log_2\frac{25}{60} + \frac{35}{60}\log_2\frac{35}{60}) = 0.9799$

   For Split 1:

$$\begin{aligned}
&Tbl(S_1|A) = [20+, 15-] \text{ implies } H(S_1|A) = 0.9852\\
&Tbl(S_1|B) = [5+, 20-] \text{ implies } H(S_1|B) = 0.7219\\
&IG(S_1) = 0.9799 - \left[\frac{35}{60}(0.9852) + \frac{25}{60}(0.7219)\right] = 0.1044\\
&I(S_1) = -\left[\frac{35}{60}\log_2\left(\frac{35}{60}\right) + \frac{25}{60}\log_2\left(\frac{25}{60}\right)\right] = 0.9799\\
&R(S_1) = \frac{0.1044}{0.9799} = 0.1065
\end{aligned}$$

   For Split 2:

$$\begin{aligned}
&Tbl(S_2|A) = [10+, 2-] \text{ implies } H(S_2|A) = 0.6500\\
&Tbl(S_2|B) = [15+, 33-] \text{ implies } H(S_2|B) = 0.8960\\
&IG(S_2) = 0.9799 - \left[\frac{12}{60}(0.6500) + \frac{48}{60}(0.8960)\right] = 0.1331\\
&I(S_2) = -\left[\frac{12}{60}\log_2\left(\frac{12}{60}\right) + \frac{48}{60}\log_2\left(\frac{48}{60}\right)\right] = 0.7219\\
&R(S_2) = \frac{0.1331}{0.7219} = 0.1844
\end{aligned}$$

   Split 2 has a higher gain ratio, hence Split 2 is preferred. □

2. (Jan 2022 Final Q4(b)) A classification tree is being constructed to predict whether the credit card application approval is positive. Consider the two splits below:

   - **Split 1**: The left node has 178 observations with 68 positives and the right node has 295 observations with 144 positives.
   - **Split 2**: The left node has 136 observations with 83 positives and the right node has 337 observations with 129 positives.

   By calculating the information gains, determine which split is better. (7 marks)

   Solution. First, we calculate the entropy of Y:

$$H(Y) = -\left(\frac{212}{473}\log_2\frac{212}{473} + \frac{261}{473}\log_2\frac{261}{473}\right) = 0.9922448 \qquad \text{[2 marks]}$$

The entropy of **Split 1** is

$$H_1 = \frac{178}{473}\left[-\frac{68}{178}\log_2\frac{68}{178} - \frac{110}{178}\log_2\frac{110}{178}\right] + \frac{295}{473}\left[-\frac{144}{295}\log_2\frac{144}{295} - \frac{151}{295}\log_2\frac{151}{295}\right]$$
$$= 0.9844898$$

[1.5 marks]

The information gain for **Split 1** is

$$IG_1 = H(Y) - H_1 = 0.007755 \qquad [0.5\text{ mark}]$$

The entropy of **Split 2** is

$$H_2 = \frac{136}{473}\left[-\frac{83}{136}\log_2\frac{83}{136} - \frac{53}{136}\log_2\frac{53}{136}\right] + \frac{337}{473}\left[-\frac{129}{337}\log_2\frac{129}{337} - \frac{208}{337}\log_2\frac{208}{337}\right]$$
$$= 0.961317$$

[1.5 marks]

The information gain for **Split 2** is

$$IG_2 = H(Y) - H_2 = 0.0309278 \qquad [0.5\text{ mark}]$$

Split 2: One node has 10 observations with 8 lapses and one node has 90 observations with 22 lapses.

$$H(S) = -\left[\frac{30}{100}\log_2\frac{30}{100} + \frac{70}{100}\log_2\frac{70}{100}\right] = 0.8813$$
$$H(Split2) = -\frac{10}{100}\left[\frac{8}{10}\log_2\frac{2}{10} + \frac{2}{10}\log_2\frac{2}{10}\right] - \frac{90}{100}\left[\frac{22}{90}\log_2\frac{22}{90} + \frac{68}{90}\log_2\frac{68}{90}\right]$$
$$= -0.1 \times (-0.7219) - 0.9 \times (-0.8024) = 0.7943$$
$$IG(Split2) = 0.8813 - 0.7943 = 0.0870$$

Since **Split 2** has a higher information gain, it is a better split. .... [1 mark]

□

3. (May 2020 Final Q4(b)(ii)) In trying to build a model that is able to predict whether or not an email message is spam based on the following predictors:

- to_multiple: Indicator for whether the email was addressed to more than one recipient;
- image: Indicates whether any images were attached;
- attach: Indicates whether any files were attached;
- dollar: Indicates whether a dollar sign or the word 'dollar' or 'ringgit' appeared in the email;
- winner: Indicates whether "winner" appeared in the email;
- num_char: The number of characters in the email, in thousands;
- format: Indicates whether the email was written using HTML (e.g. may have included bolding or active links) or plaintext;
- re_subj: Indicates whether the subject started with "Re:", "RE:", "re:", or "rE:";
- number: Factor variable saying whether there was no number, a small number (under 1 million), or a big number.

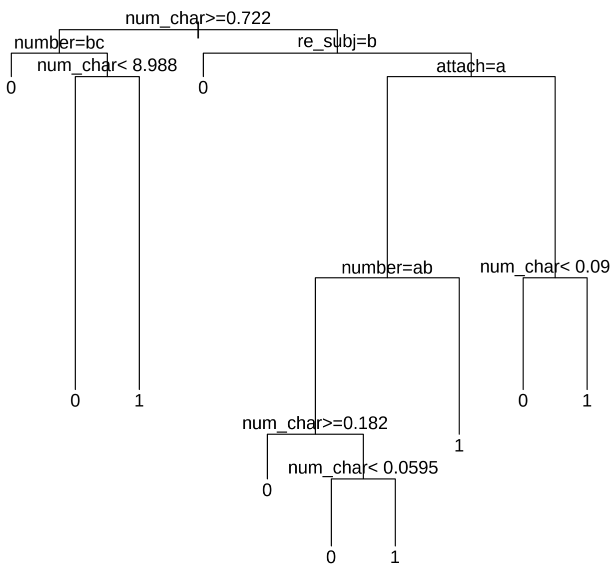
Note that "spam" is denoted with the value 1 while "non-spam" is denoted with the value 0. The trained logistic regression model has the parameters given in Figure 4.2.

Table 4.2: Coefficients of Logistic Regression

```
Coefficients:
                 Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)     -1.468478   0.181285   -8.100  5.48e-16 ***
to_multipleyes  -2.152057   0.349538   -6.157  7.42e-10 ***
imageyes        -1.467843   0.797895   -1.840  0.065820 .
attachyes        0.957716   0.281455    3.403  0.000667 ***
num_char        -0.014651   0.007199   -2.035  0.041849 *
dollaryes        0.453477   0.197009    2.302  0.021346 *
winneryes        1.994563   0.392252    5.085  3.68e-07 ***
numbersmall     -1.227981   0.186300   -6.591  4.36e-11 ***
numberbig       -0.561313   0.263563   -2.130  0.033195 *
formatPlain      1.032511   0.171915    6.006  1.90e-09 ***
re_subjyes      -2.447223   0.398309   -6.144  8.05e-10 ***
---
Signif. :   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If an email does not address to multiple, has no image, no attached file(s), no "dollar" sign, does not have the word "winner", has 20.133×10^3 number of characters and is in HTML format, has no subject starting with "Re:" and has a small number in the email. **Determine** whether the email is a spam using the trained logistic regression model and using the decision tree model (you will need to interpret the decision tree model based on your knowledge of "rpart" algorithm) given in Figure 4.3.

Figure 4.3: The trained decision tree model.



(4.5 marks)

Solution. Given the predictors, the probability of spam is

$$\mathbb{P}(Y=1|X=x) = \frac{1}{1+\exp(-(-1.468478+\beta^T x))} \qquad [1\text{ mark}]$$
$$= \frac{1}{1+\exp(1.468478+1.52295)} = 0.047815$$

where

$$\beta^T x = -2.152057*0 - 1.467843*0 + 0.957716*0 - 0.014651*20.133$$
$$+ 0.453477*0 + 1.994563*0 - 1.227981 + 1.032511*0 \qquad [1.5\text{ marks}]$$
$$- 2.447223*0 = -1.52295$$

Since the probability of the email being spam is smaller than 0.5, we predict it to be a **non-spam**. . . . . [0.5 mark]

The prediction with tree is as follows:

(a) number of characters = 20.133 > 0.722, go to left subtree;

(b) number = small. It should match "b", therefore, go to left subtree, given us 0, i.e. **non-spam**.

Marks are deducted if no justification is given. . . . . [1.5 marks] ☐

4. (Jan 2021 Final Q2(a)) The dataset in Table 2.1 is used to build a classification tree which predicts if a student pass predictive modelling (Pass or Fail, P, F for short), based on their previous GPA (High, Medium, or Low, H, M, L for short) and whether they have or have not (Y or N in short) put in significant efforts in their study.

Table 2.1: Training dataset for classification problem.

GPA	Studied	Pass
L	N	F
L	Y	P
M	N	F
M	Y	P
H	N	P
H	Y	P

Construct and plot the ID3 classification tree (using information gain) with appropriate labels. You must show all the calculation steps. (5 marks)

Solution. First, we calculate the entropy

$$H = -(\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6}) = 0.9182958 \qquad [1 \text{ mark}]$$

The information gain for the variable GPA is

$$\begin{aligned} IG_1 = H - (&-\frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) \\ &- \frac{1}{3}(\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2})) \qquad [1 \text{ mark}] \\ = 0.9182958 &- \frac{1}{3}(1+1+0) = 0.2516291 \end{aligned}$$

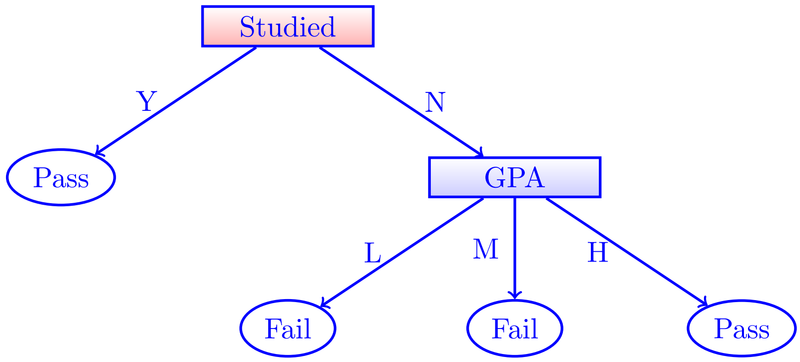
The information gain for the variable Studied is

$$\begin{aligned} IG_2 &= H - \left(-\frac{1}{2}(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) - \frac{1}{2}(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3})\right) \qquad [1 \text{ mark}] \\ &= 0.9182958 - \frac{1}{2}(0.9182958 + 0) = 0.4591479 \end{aligned}$$

The variable "Studied" is choice for the ID3 split because it has higher information gain.

For "Studied=N", we have three more branches because the output variable Pass is not pure.

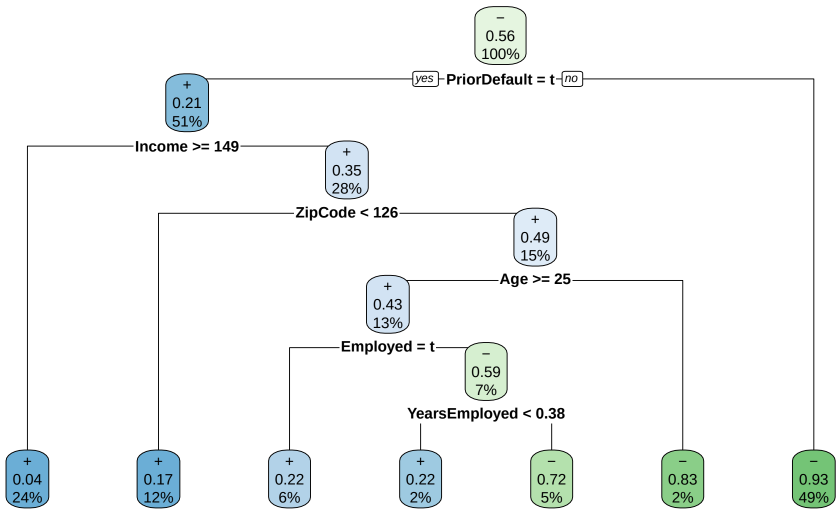
For "Studied=Y", the output variable Pass is already pure. [0.5 mark]



. . . . [1.5 marks] □

5. (Jan 2022 Final Q2(b)) For the same training data (as Tutorial 4 Q1, i.e. Jan 2022 Final Q2(b)), use the CART tree in Figure 2.1 to predict the the credit card application being approved (positive or negative) for a male of age 22.08 with a debt of 0.83 unit who has been employed for 2.165 years with no prior default and is currently unemployed, has a credit score 0 and a zip code 128 with income 0.

Figure 2.1: CART tree for credit card application approval data.



You need to show your workings by explaining the steps to move left or right in the tree travesal to reach the prediction. (4 marks)

Solution. From the decision tree, we move right (no prior default) directly to negative (Probability=0.93. It consists of 49% of the training data). . . . . [3 marks]

The credit card application being approved is negative. . . . . [1 mark] □

6. (Jan 2022 Final Q2(c)) Compare the ability of the logistic regression model and the C4.5 tree model in the handling missing values and the prediction of highly nonlinear data. (4 marks)

Solution. Logistic regression model will omit data with missing values since the mathematical model does not allow the arithmetic calculation for missing value. . . . . [1 mark]

C4.5 tree model will ignore the missing value in the feature and compute the gain ratio. [1 mark]

Logistic regression model performs poorly when it is used to predict highly nonlinear data since the model is linear. . . . . [1 mark]

C4.5 tree model performs better compare to logistic regression model when it is used to predict highly nonlinear data since it is nonlinear. However, it may overfit and does not generalise well. ....[1 mark] □

7. (Final Exam Jan 2023, Q4(a)) Consider a marketing data in Table 4.1 with `Gender`, `Car Type`, and `Cloth Size` as predictors which are categorical.

ID	Gender	Car Type	Cloth Size	Label
1	Male	B	S	−
2	Male	C	M	−
3	Male	C	M	−
4	Male	C	L	−
5	Male	C	XL	−
6	Male	C	XL	−
7	Female	C	S	−
8	Female	C	S	−
9	Female	C	M	−
10	Female	A	L	−
11	Male	B	L	+
12	Male	B	XL	+
13	Male	B	M	+
14	Male	A	XL	+
15	Female	A	S	+
16	Female	A	S	+
17	Female	A	M	+
18	Female	A	M	+
19	Female	A	M	+
20	Female	A	L	+

Table 4.1: Marketing data.

Perform computations for the impurity measurements and the decision tree construction below by using the multi-way split.

(i) Compute the Gini index for the `Gender` attribute. (3 marks)

Solution.

$$Gini(\texttt{Gender}) = \frac{10}{20}(1-(\frac{6}{10})^2-(\frac{4}{10})^2)+\frac{10}{20}(1-(\frac{4}{10})^2-(\frac{6}{10})^2) = 0.48 \quad \text{[3 marks]}$$

Average: 1.66 / 3 marks in Jan 2023; 12% below 1.5 marks.

□

(ii) Compute the Gini index for the `Car Type` attribute. (4 marks)

Solution.

$$\begin{aligned} &Gini(\texttt{Car Type}) \\ =&\frac{8}{20}(1-(\frac{1}{8})^2-(\frac{7}{8})^2)+\frac{4}{20}(1-(\frac{1}{4})^2-(\frac{3}{4})^2)+\frac{8}{20}(1-(\frac{8}{8})^2-(\frac{0}{8})^2) \quad \text{[4 marks]} \\ =&0.1625 \end{aligned}$$

Average: 2.05 / 4 marks in Jan 2023; 12% below 2 marks. □

(iii) Compute the Gini index for the `Cloth Size` attribute. (4 marks)

Solution.

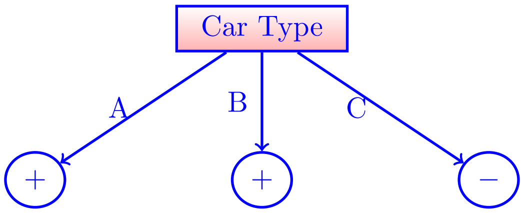
$$
\begin{aligned}
&Gini(\texttt{Cloth Size}) \\
=&\frac{5}{20}(1-(\frac{3}{5})^2-(\frac{2}{5})^2)+\frac{7}{20}(1-(\frac{3}{7})^2-(\frac{4}{7})^2) \\
&+\frac{4}{20}(1-(\frac{2}{4})^2-(\frac{2}{4})^2)+\frac{4}{20}(1-(\frac{2}{4})^2-(\frac{2}{4})^2) \\
=&0.4914286
\end{aligned}
$$

[4 marks]

Average: 2.05 / 4 marks in Jan 2023; 12% below 2 marks. □

(iv) Construct a multiway-split decision tree with only one level based on the results from part (i) to part (iii) above. (2 marks)

Solution. Based on the result from part (i) to part (iii), Car Type is the most suitable attribute and a one-level multiway decision tree is



. . . . [2 marks]

Average: 0.29 / 2 marks in Jan 2023; 21% below 1 mark. □

(v) Describe the workings of the random forest predictive model given the $n \times (p+1)$ data D. (3 marks)

Solution.

- Sample $m < p$ columns/features randomly from the $n \times (p+1)$ data D with replacement as data D_t. . . . . [1 mark]
- Grow a simple decision tree of level 1 or a CART tree for the bootstrap data D_t [1 mark]
- Stop when T number of decision trees are constructed. The random forest is the collection of decision trees. . . . . [1 mark]

Average: 0.37 / 3 marks in Jan 2023; 25% below 1.5 marks. □

8. (Final Exam May 2023, Q2(b))

(a) Given the following R output of the bank customer churn dataset:

```
        Exited
Age<42.5    0    1
   FALSE  833  604
   TRUE  3148  414
                 Exited
NumOfProducts<2.5    0    1
            FALSE   24  140
            TRUE  3957  878
              Exited
IsActiveMember    0    1
             0 1779  648
             1 2202  370
```

Compute the Gini index for the `Age` using the cutoff 42.5, the Gini index for the `NumOfProducts` using the cutoff 2.5 and the Gini index for `IsActiveMember` and determine which one of them is the best attribute for the root of a C4.5 decision tree. (9 marks)

Solution.

$$\begin{aligned} G(Age < 42.5) &= \frac{833+604}{4999}\left(1-\left(\frac{833}{833+604=1437}\right)^2-\left(\frac{604}{1437}\right)^2\right) \\ &\quad+\frac{3148+414}{4999}\left(1-\left(\frac{3148}{3148+414=3562}\right)^2-\left(\frac{414}{3562}\right)^2\right) \\ &= 0.286461 \end{aligned}$$

[3 marks]

$$\begin{aligned} G(NumOfProducts < 2.5) & \\ &= \frac{24+140}{4999}\left(1-\left(\frac{24}{24+140=164}\right)^2-\left(\frac{140}{164}\right)^2\right) \\ &\quad+\frac{3957+878}{4999}\left(1-\left(\frac{3957}{3957+878=4835}\right)^2-\left(\frac{878}{4835}\right)^2\right) \\ &= 0.295679 \end{aligned}$$

[3 marks]

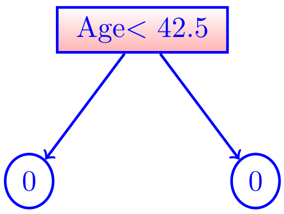
$$\begin{aligned} G(IsActiveMember) & \\ &= \frac{1779+648}{4999}\left(1-\left(\frac{1779}{1779+648=2427}\right)^2-\left(\frac{648}{2427}\right)^2\right) \\ &\quad+\frac{2202+370}{4999}\left(1-\left(\frac{2202}{2202+370=2572}\right)^2-\left(\frac{370}{2572}\right)^2\right) \\ &= 0.316767 \end{aligned}$$

[2 marks]

Since Age< 42.5 has the lowest Gini impurity index, it is the best attribute for the root of a C4.5 decision tree. ....[1 mark] □

(b) Construct a decision tree with only one level based on the Gini index results from part (i). (2 marks)

Solution. Based on the result from part (i).



..................................[2 marks] □

9. (Final Exam May 2023, Q5(a)) Given the training data with features X_1, X_2 and the label Y in Table 5.1.

Obs.	Petal.Length	Petal.Width	Sepal.Length	Species
1	1.5	0.2	5.0	setosa
2	1.1	0.1	4.3	setosa
3	4.0	1.2	5.8	versicolor
4	3.3	1.0	4.9	versicolor
5	5.4	2.1	6.9	virginica
6	5.1	1.9	5.8	virginica

Table 5.1: Training data with features Petal.Length, Petal.Width, Sepal.Length and the label Species of iris flower.

(i) A decision tree is trained on the training data from Table 5.1 and is shown in Figure 5.1.

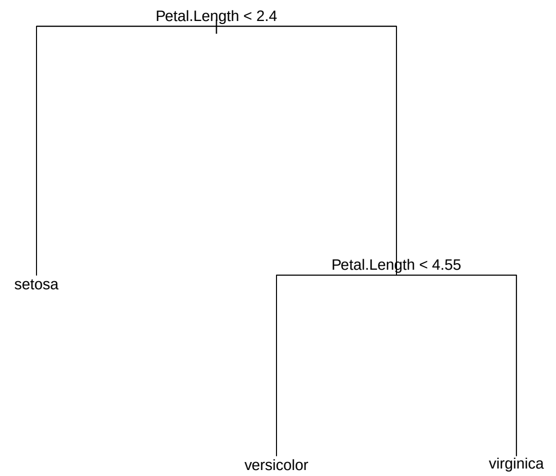


Figure 5.1: Tree predictive model trained on data from Table 5.1.

Use the decision tree to predict the species of the iris flower with a petal length of 3.9, a petal width of 1.4 and a sepal length of 5.2. (4 marks)

Solution. The petal length of 3.9 is more than 2.4, **go to the right subtree**. [2 marks] Then petal length is less than 4.55, **go to the left subtree which reaches the species versicolor**. . . . . [2 marks] □

(ii) State the reason for a trained decision tree to be more efficient in prediction than a kNN for data when the number of samples, n, is large, from a computational point of view. (2 marks)

Solution. The reason for decision tree to be more effience is because the tree partitions the data into multiple segments leading to a tree with a depth of mostly $\log_2 n$ where as kNN needs to compare the distance of the input to all n training data in order to perform prediction. . . . . [2 marks] □

11. (Final Exam Jan 2024 Sem, Q2) When a bank receives a loan application, the bank has to make a decision whether to go ahead with the loan approval or not based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank;
- If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank.

To minimise loss from the bank's perspective, the bank needs a predictive model regarding who to give approval of the loan and who not to based on an applicant's demographic and socio-economic profiles.

Suppose the response variable `Y` is 0 when the loan is approved and is 1 when the loan is not approved. Suppose the features of the data are listed below:

- X_1 (categorical): Status of existing checking account (A11, A12, A13, A14);
- X_2 (integer): Duration in months
- X_3 (integer): Credit amount
- X_4 (integer): Instalment rate in percentage of disposable income
- X_5 (binary): foreign worker (yes, no)

(c) When the data is trained with a CART model the text representation of the CART is obtained:

```
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 80 105.900 0 ( 0.6250 0.3750 )
   2) X1: A13,A14 38  33.150 0 ( 0.8421 0.1579 )
     4) X4 < 2.5 12    0.000 0 ( 1.0000 0.0000 ) *
     5) X4 > 2.5 26   28.090 0 ( 0.7692 0.2308 )
      10) X2 < 30 20   16.910 0 ( 0.8500 0.1500 )
        20) X3 < 1550.5 10   12.220 0 ( 0.7000 0.3000 ) *
        21) X3 > 1550.5 10    0.000 0 ( 1.0000 0.0000 ) *
      11) X2 > 30 6    8.318 0 ( 0.5000 0.5000 ) *
   3) X1: A11,A12 42   57.360 1 ( 0.4286 0.5714 )
     6) X3 < 3266.5 29   40.170 0 ( 0.5172 0.4828 )
      12) X3 < 1499 16   19.870 1 ( 0.3125 0.6875 )
        24) X4 < 2.5 5    0.000 1 ( 0.0000 1.0000 ) *
        25) X4 > 2.5 11   15.160 1 ( 0.4545 0.5455 ) *
      13) X3 > 1499 13   14.050 0 ( 0.7692 0.2308 )
        26) X3 < 2243.5 7    0.000 0 ( 1.0000 0.0000 ) *
        27) X3 > 2243.5 6    8.318 0 ( 0.5000 0.5000 ) *
     7) X3 > 3266.5 13   14.050 1 ( 0.2308 0.7692 )
      14) X3 < 6595.5 8    0.000 1 ( 0.0000 1.0000 ) *
      15) X3 > 6595.5 5    6.730 0 ( 0.6000 0.4000 ) *
```

Apply the CART model to predict Y for a foreign worker when the status of existing checking account of the customer is A11, the duration is 6 months, the credit amount is 1169 and the instalment rate of disposable income is 4%. You need to write down your steps. (3 marks)

Solution.
- $X_1 = A11$, go to item 3) . . . . [0.5 mark]
- $X_3 = 1169 < 3266.5$, go to 6) . . . . [0.5 mark]
- $X_3 = 1169 < 1499$, go to 12) . . . . [0.5 mark]
- $X_4 = 4 > 2.5$, $\boxed{Y = 1}$. . . . . [1.5 marks]

Average: 1.17 / 3 marks in Jan 2024; 58.18% below 1.5 marks.

Remark: I mentioned during practical that although programming questions will not come out, the computer outputs from R analysis needs to be recognised. However, most students are doing other projects and other assignments during practical class leading to a poor result. □

(d) Suppose the confusion matrix for logistic regression is given in Table 2.1, the confusion matrix for naive Bayes model is given in Table 2.2, the confusion matrix for CART model is given in Table 2.3, if your objective is to identify the applicant with good credit risk and reject applicants with bad credit risk, state the performance metrics that meets your requirement and evaluate if the models are acceptable based on appropriate performance metrics calculations.

Table 2.1: Confusion matrix for Logistic Regression (0 is positive)

	Actual	
Prediction	0	1
0	466	98
1	184	172

Table 2.2: Confusion matrix for naive Bayes model (0 is positive)

	Actual	
Prediction	0	1
0	556	174
1	94	96

Table 2.3: Confusion matrix for CART model (0 is positive)

	Actual	
Prediction	0	1
0	446	142
1	204	128

(4 marks)

Solution. Since the data are **imbalanced** (650 zeros vs 270 ones), accuracy is not a good performance metric:

- Accuracy of logistic regression = 0.6934783
- Accuracy of naive Bayes model = 0.7086957
- Accuracy of CART model = 0.623913

None of the three models are acceptable because if we predict all to be zeros, we get an accuracy of $650/(650 + 270) = 0.7065217$. .... [3 marks]

A better performance metric is the Kappa statistic which captures the recalls and the precision. .... [1 mark]

Average: 1.42 / 3 marks in Jan 2024; 52.73% below 1.5 marks. □