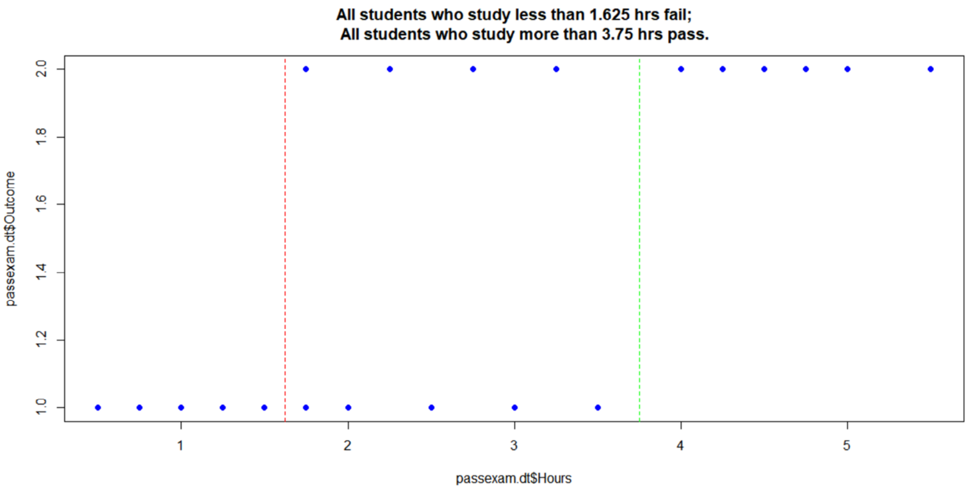
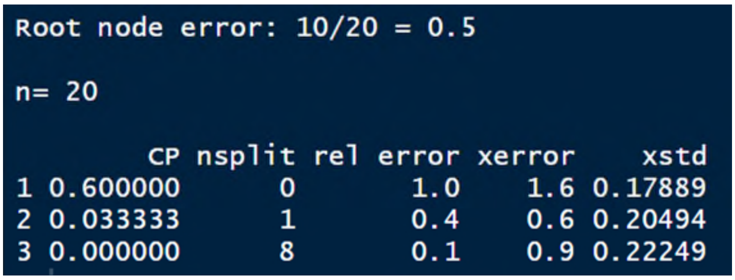
Exercise 8.2: CART Analysis for passexam.csv and passexam2.csv

1. In previous unit on Logistic Regression, we analyzed passexam.csv. Plotting the dataset shows at least two clear cut-offs for failing and passing the exam. Can CART detect the two cut-offs?



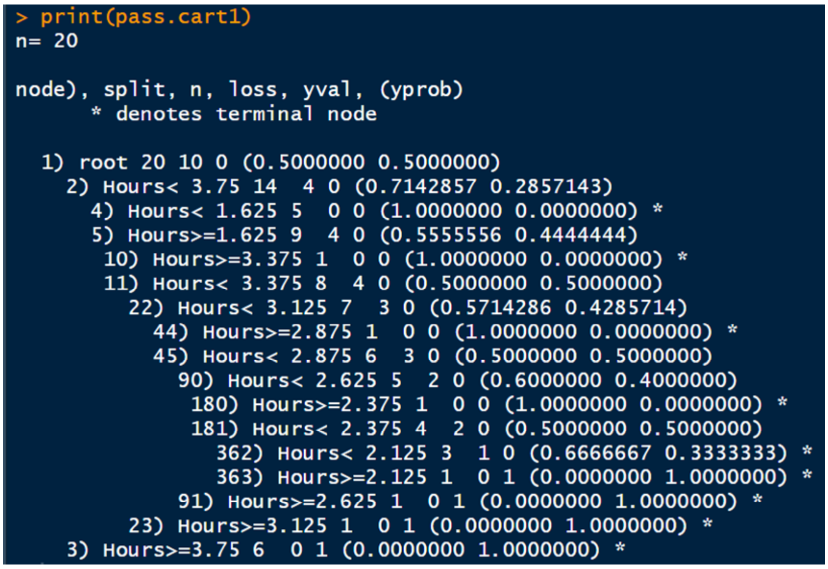
Solution:



```
Root node error: 10/20 = 0.5

n= 20

        CP nsplit rel error xerror    xstd
1 0.600000      0       1.0    1.6 0.17889
2 0.033333      1       0.4    0.6 0.20494
3 0.000000      8       0.1    0.9 0.22249
```

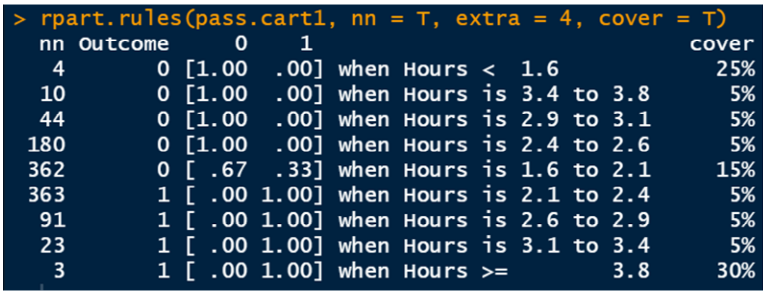
CP table reveals either 1 split or 8 splits (maximal tree).

Printing the maximal tree shows that node 3 and node 4 correctly identifies the two cutoffs.



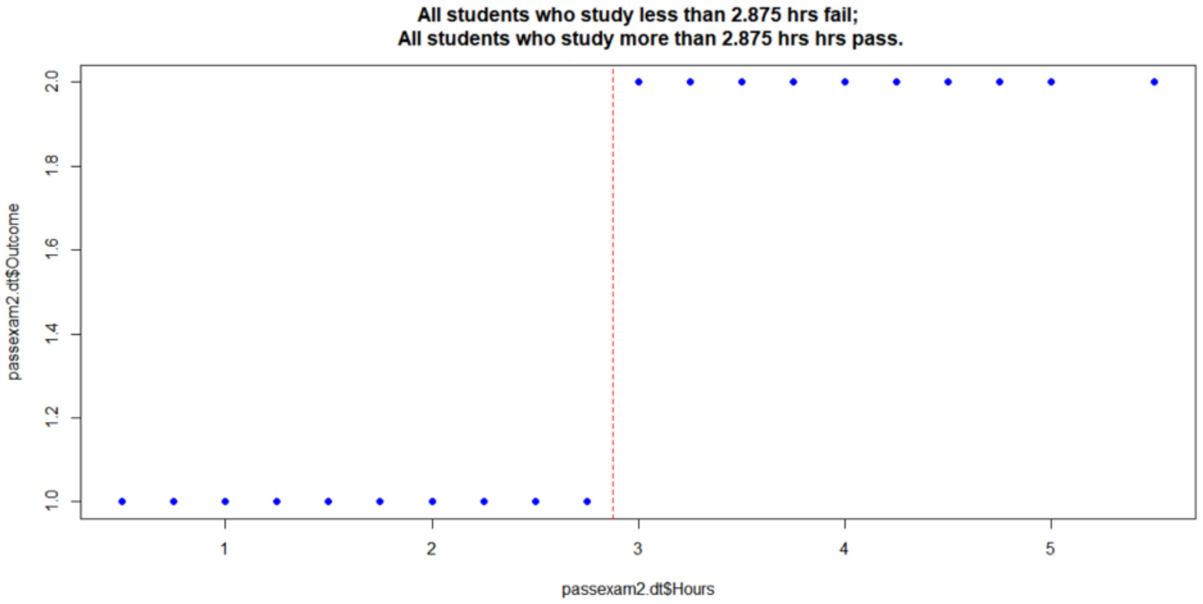
```
> print(pass.cart1)
n= 20

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 20 10 0 (0.5000000 0.5000000)
    2) Hours< 3.75 14  4 0 (0.7142857 0.2857143)
      4) Hours< 1.625 5  0 0 (1.0000000 0.0000000) *
      5) Hours>=1.625 9  4 0 (0.5555556 0.4444444)
       10) Hours>=3.375 1  0 0 (1.0000000 0.0000000) *
       11) Hours< 3.375 8  4 0 (0.5000000 0.5000000)
         22) Hours< 3.125 7  3 0 (0.5714286 0.4285714)
           44) Hours>=2.875 1  0 0 (1.0000000 0.0000000) *
           45) Hours< 2.875 6  3 0 (0.5000000 0.5000000)
             90) Hours< 2.625 5  2 0 (0.6000000 0.4000000)
              180) Hours>=2.375 1  0 0 (1.0000000 0.0000000) *
              181) Hours< 2.375 4  2 0 (0.5000000 0.5000000)
                362) Hours< 2.125 3  1 0 (0.6666667 0.3333333) *
                363) Hours>=2.125 1  0 1 (0.0000000 1.0000000) *
             91) Hours>=2.625 1  0 1 (0.0000000 1.0000000) *
         23) Hours>=3.125 1  0 1 (0.0000000 1.0000000) *
    3) Hours>=3.75 6  0 1 (0.0000000 1.0000000) *
```

The rules found can be reconciled and simplified by rpart.rules() function.



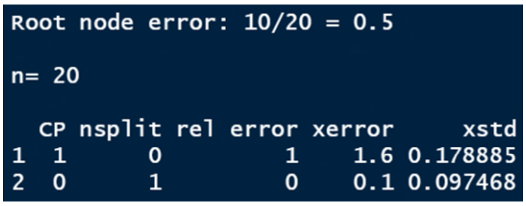
```
> rpart.rules(pass.cart1, nn = T, extra = 4, cover = T)
  nn Outcome      0    1                             cover
   4        0 [1.00  .00] when Hours <  1.6            25%
  10        0 [1.00  .00] when Hours is 3.4 to 3.8      5%
  44        0 [1.00  .00] when Hours is 2.9 to 3.1      5%
 180        0 [1.00  .00] when Hours is 2.4 to 2.6      5%
 362        0 [ .67  .33] when Hours is 1.6 to 2.1     15%
 363        1 [ .00 1.00] when Hours is 2.1 to 2.4      5%
  91        1 [ .00 1.00] when Hours is 2.6 to 2.9      5%
  23        1 [ .00 1.00] when Hours is 3.1 to 3.4      5%
   3        1 [ .00 1.00] when Hours >=        3.8     30%
```

2. Previously, we learnt that Logistic Regression fails on perfectly separable data passexam2.csv. Plotting the dataset shows at least one clear cut-off for failing and passing the exam. Can CART succeed in creating the model and detect the single cut-off?

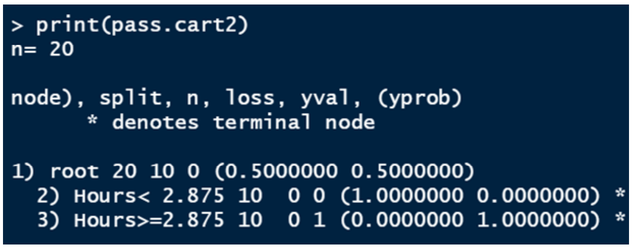


Solution:

CART succeeds in finding the cut-off 2.875 hrs.



```
Root node error: 10/20 = 0.5

n= 20

  CP nsplit rel error xerror     xstd
1  1      0         1    1.6 0.178885
2  0      1         0    0.1 0.097468
```



```
> print(pass.cart2)
n= 20

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 20 10 0 (0.5000000 0.5000000)
  2) Hours< 2.875 10  0 0 (1.0000000 0.0000000) *
  3) Hours>=2.875 10  0 1 (0.0000000 1.0000000) *
```