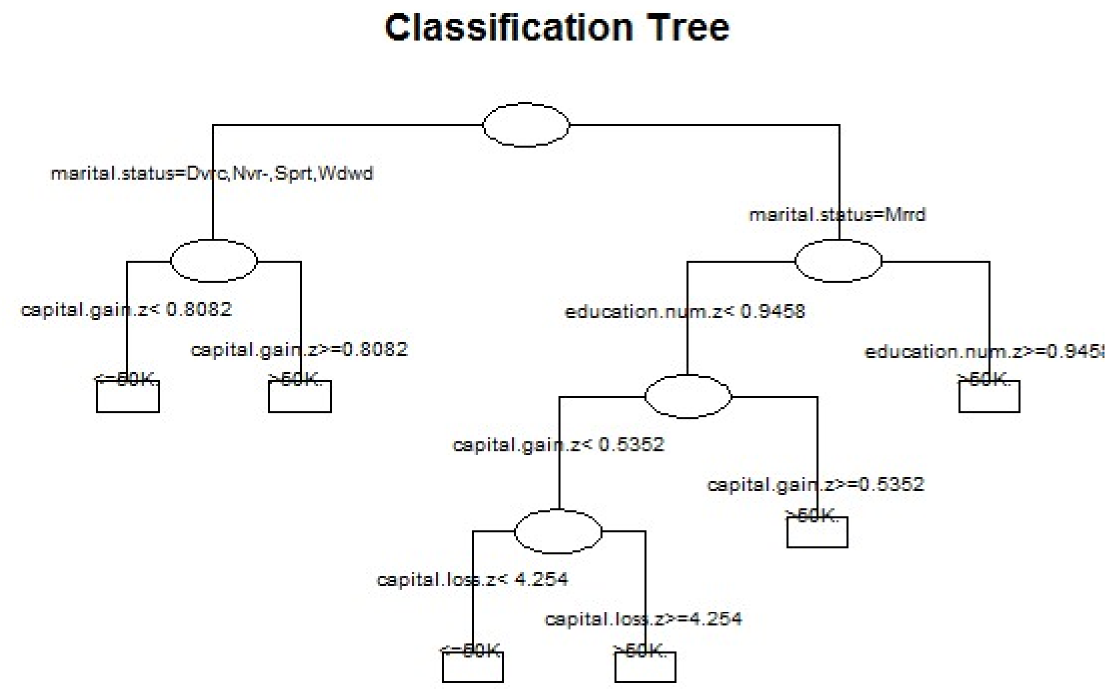
Problem set 6

Lina Mi

Problem 1: Follow the instructions of R zone, the decision tree generated using CART algorithm for example dataset (problem_set-6-Clem3Training.csv) is shown as following



the output result of `print(cartfit)` is as following

```
n= 25000 node), split, n, loss, yval, (yprob)

     * denotes terminal node

 1) root 25000 5984 <=50K. (0.76064000 0.23936000)

   2) marital.status=Divorced,Never-married,Separated,Widowed 13215  845 <=50K. (0.93605751 0.06394249)

     4) capital.gain.z< 0.8082312 12986  625 <=50K. (0.95187125 0.04812875) *

     5) capital.gain.z>=0.8082312 229    9 >50K. (0.03930131 0.96069869) *

   3) marital.status=Married 11785 5139 <=50K. (0.56393721 0.43606279)

     6) education.num.z< 0.9458454 8296 2672 <=50K. (0.67791707 0.32208293)

      12) capital.gain.z< 0.5352109 7894 2280 <=50K. (0.71117304 0.28882696)
     24) capital.loss.z< 4.254168 7615 2076 <=50K. (0.72738017 0.27261983) *
     25) capital.loss.z>=4.254168 279   75 >50K. (0.26881720 0.73118280) *
   13) capital.gain.z>=0.5352109 402   10 >50K. (0.02487562 0.97512438) *
  7) education.num.z>=0.9458454 3489 1022 >50K. (0.29292061 0.70707939) *
```

From the decision tree and summary of CART, it can be seen that

1) root 25000 5984 <=50K. (0.76064000 0.23936000)
   The root node represent all observations(25000), the whole dataset is classified as <=50k (target variable: income), with 5984 will be incorrectly classified as >50K, with 76.06% of observations have the target variable **income** as <=50K and 23.94% of observations have it as >50K. the root node is split into two sub-nodes, nodes (2)and (3), based on the variable marrital.status with a split value of Married or other marital status.

2) marital.status=Divorced,Never-married,Separated,Widowed 13215 845 <=50K. (0.93605751 0.06394249)
   this node with observations of 13215, with target variable <=50K, 845 of these 13215 observations are misclassified, representing 93.61% of accuracy. This node is further split into two leaf nodes (4) and (5) based on the variable **capital.gain** with split value 0.8082312;

   4) **capital.gain.z**< 0.8082312 12986 625 <=50K. (0.95187125 0.04812875) *
   The observations of 12986 with **variable capital.gain.z**<0.81 and variable **marital.status**=Divorced, Never-married, Seperated, Widowed are classified as income <=50K, the accuracy if 93. 61%

   5) capital.gain.z>=0.8082312 229   9 >50K. (0.03930131 0.96069869) *
   The 229 observations with variable capital.gain.z>=0.8082312 and variable **marital.status**= Divorced, Never-married, Seperated, Widowed are classified as income >50K, 9 out of 229 are misclassified and accuracy of this leaf node is 96.07%

3) marital.status=Married 11785 5139 <=50K. (0.56393721 0.43606279)
   this node with 11785 observations with variable **marital.status**=Married, all of observations are classified as **income**<=50K, and 5139 of 11785 observations are misclassified, with close to 50/50 split in this partition, this node is further partitioned into one sub-node (6) and one leaf node (7), based on the variable **education.num.z** with split value 0.9458454

   6)**education.num.z**< 0.9458454 8296 2672 <=50K. (0.67791707 0.32208293)
   this subset are consist of observations with variable **marital.status**=Married and **education.num.z**< 0.9458454, target variable are classified as **income**<=50K, 2672 out of 8296 are misclassifed and accuracy is 67.8%. this node is further partitioned into one subnode (12) and one leaf node(13) , according to criteria of variable **capital.gain.z** >= 0.5352109 or <0.5352109

   12) **capital.gain.z**< 0.5352109 7894 2280 <=50K. (0.71117304 0.28882696)
   This subset of observations are classified as target variable income<=50 with accuracy of 71.11%, this node is further partitioned into two leaf nodes (24) and (25), based on criteria of variable **capital.loss.z**< 4.254168 or >= 4.254168

24) capital.loss.z< 4.254168 7615 2076 <=50K. (0.72738017 0.27261983)

This leaf node are consist of observations with marital.status=Married, **education.num.z**< 0.9458454, **capital.gain.z**< 0.5352109 capital.loss.z< 4.254168, this leaf node has target variable of income <=50K, accuracy of it 72.74%

25) capital.loss.z>=4.254168 279 75 >50K. (0.26881720 0.73118280)

This leaf node are consist of observations with marital.status=Married, **education.num.z**< 0.9458454, **capital.gain.z**< 0.5352109 capital.loss.z >=4.254168, this leaf node has target variable of income >50K, accuracy of it 73.12%

13) **capital.gain.z**>=0.5352109 402 10 >50K. (0.02487562 0.97512438)

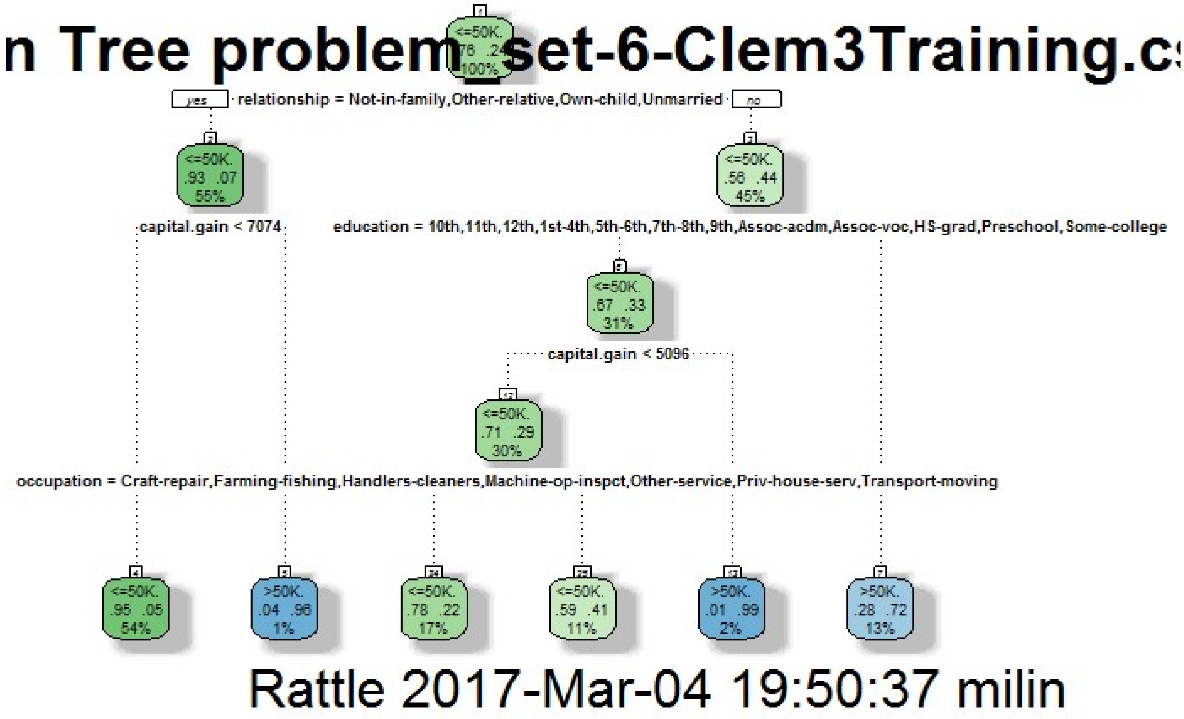
This node is leaf node with 402 observations in the sub-dataset, with variable marital.status=Married, **education.num.z**< 0.9458454, **capital.gain.z**>= 0.5352109, this node is classifyed as income >50K with accuracy of 97.51%

7) **education.num.z**>=0.9458454 3489 1022 >50K. (0.29292061 0.70707939)

this is leaf node, consisting of observations with variable **marital.status**=Married and **education.num.z**>= 0.9458454, target variable are classified as **income**>50K, accuracy is 70.71%.

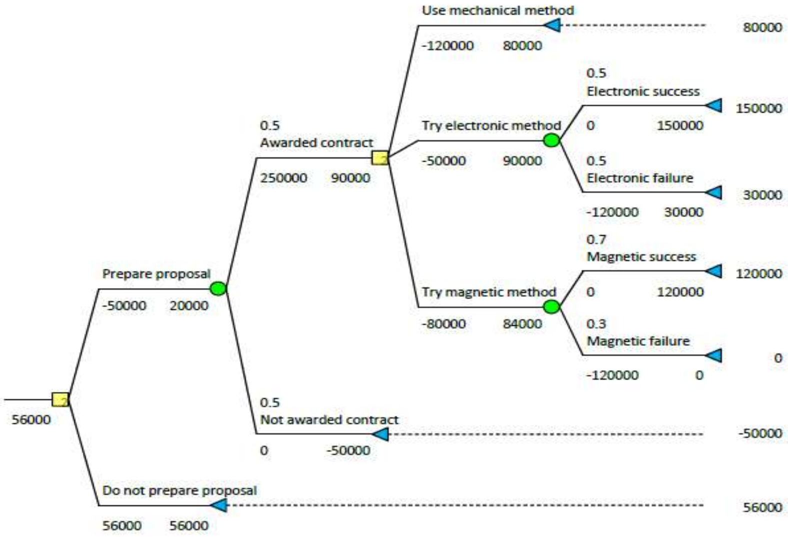
Problem2: Rattle Decision Tree:

Using Rattle and rpart, I got the decision tree on same example dataset and the decision tree is shown as following:


n Tree problem_set-6-Clem3Training.cs
<=50K.
.76 .24
100%
yes
relationship = Not-in-family,Other-relative,Own-child,Unmarried
no
<=50K.
.93 .07
55%
<=50K.
.56 .44
45%
capital.gain < 7074
education = 10th,11th,12th,1st-4th,5th-6th,7th-8th,9th,Assoc-acdm,Assoc-voc,HS-grad,Preschool,Some-college
<=50K.
.67 .33
31%
capital.gain < 5096
<=50K.
.71 .29
30%
occupation = Craft-repair,Farming-fishing,Handlers-cleaners,Machine-op-inspct,Other-service,Priv-house-serv,Transport-moving
<=50K.
.95 .05
54%
>50K.
.04 .96
1%
<=50K.
.78 .22
17%
<=50K.
.59 .41
11%
>50K.
.01 .99
2%
>50K.
.28 .72
13%
Rattle 2017-Mar-04 19:50:37 milin

Comparing above decision tree with that in problem 1, it can be seen that each split of this decision tree is based on different criteria variables (1, **relationship**; 2, **education**; 3, **capital.gain.z**; 4,**occupation**) and different split value for each of criteria variables, so it can be concluded that different algorithm of decision trees have different strategies of partitioning dataset and hence resulting different decision tree.

Problem3: TreePlan Decision Tree



Comparing this treeplan with the cartfit decision tree and Rattle Decision tree, it can be seen that cartfit decision tree and Rattle Decision tree are top-down decision tree, though they have different split algorithm and strategy, the calculation are started from root node to a leaf node along a specific branch, the decision is made on each of split node based on particular strategy(algorithm), once the decision is made on one decision node, it influence the nodes below current decision node, the nodes on top of current node won't be affected. This is main character of top-down decision tree. For treeplan decision tree, the calculation is executed from leaf nodes to root node along all associated branches, it is rollback calculation. When some conditions changed (as in this case: the cost of **do not prepare proposal** changed from 0 to 56000), then tree plan do roll back calculations again, all of associated nodes with the change will be affected, comparing the decision tree with condition changed with the original decision tree, it can be seen that the branch condition change led to change of the priority of root decision node, decreasing from 1 to 2, and the root decision node has the same importance with the decision node of choosing different method to implement the project . in the bottom-up decision tree, the higher level decision nodes are affected by the branches and nodes under them.