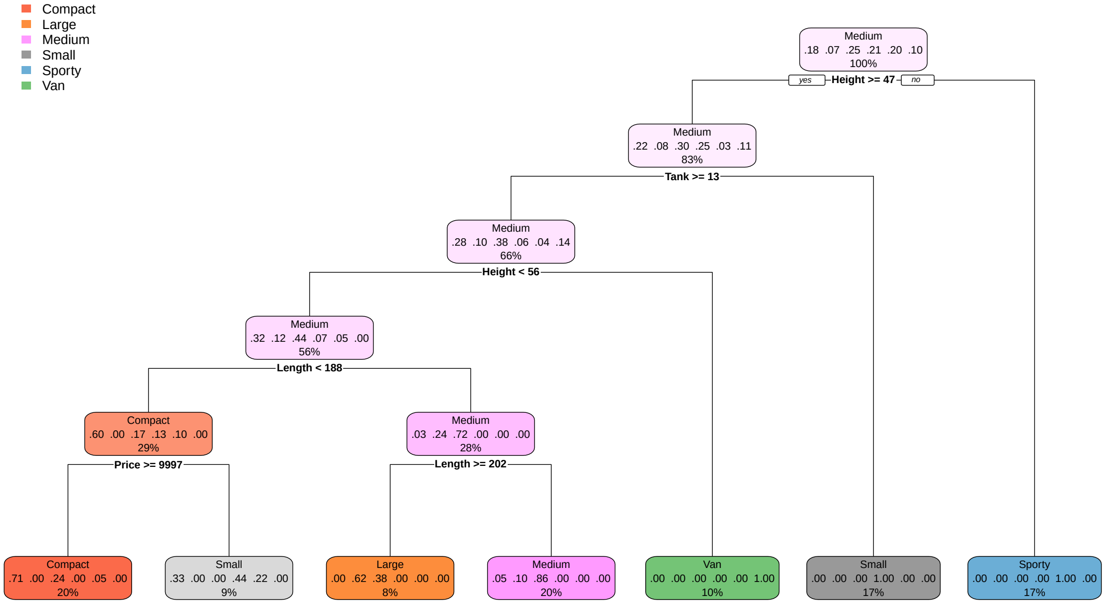
Math 154, Fall 2017, WU #5 (Trees)

Jo Hardin, Pomona College

October 30, 2017

Consider the tree below which partitions the sample of 105 cars into regions corresponding to the type of car (Compact, Large, Medium, Small, Sporty, Van).

```
library(rpart.plot); library(caret)
fitControl <- trainControl(method="none")
cars.train<- train(Type ~ Length + Height + Price + Tank, data=car90, method="rpart2",
                   na.action = na.omit, trControl = fitControl, tuneGrid= data.frame(maxdepth=10))
rpart.plot(cars.train$finalModel)
```



```
cars.train$finalModel
```

```
## n= 105
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 105 79 Medium (0.18 0.067 0.25 0.21 0.2 0.095)
##    2) Height>=47.25 87 61 Medium (0.22 0.08 0.3 0.25 0.034 0.11)
##      4) Tank>=13.4 69 43 Medium (0.28 0.1 0.38 0.058 0.043 0.14)
##        8) Height< 55.5 59 33 Medium (0.32 0.12 0.44 0.068 0.051 0)
##         16) Length< 187.5 30 12 Compact (0.6 0 0.17 0.13 0.1 0)
##           32) Price>=9997 21  6 Compact (0.71 0 0.24 0 0.048 0) *
##           33) Price< 9997 9  5 Small (0.33 0 0 0.44 0.22 0) *
##         17) Length>=187.5 29  8 Medium (0.034 0.24 0.72 0 0 0)
##           34) Length>=201.5 8  3 Large (0 0.62 0.38 0 0 0) *
##           35) Length< 201.5 21  3 Medium (0.048 0.095 0.86 0 0 0) *
##        9) Height>=55.5 10  0 Van (0 0 0 0 0 1) *
##      5) Tank< 13.4 18  0 Small (0 0 0 1 0 0) *
##    3) Height< 47.25 18  0 Sporty (0 0 0 0 1 0) *
```

1. There are $|T_0| = 7$ terminal nodes in the tree above. Find the subtree T with $|T| < 7$ terminal nodes which has the minimum overall classification error rate. (Simply prune back one step.)

$$\text{overall classification error rate} = \frac{1}{n} \sum_{m=1}^{|T|} \sum_{i \in R_m} I(y_i \neq k(m))$$

where $k(m)$ is the majority class in node m.

Answer: By removing the last node on the left (split into compact and small), there is only one additional error made. The model went from 11 errors on that branch for full tree to 12 errors on the branch for the pruned tree.

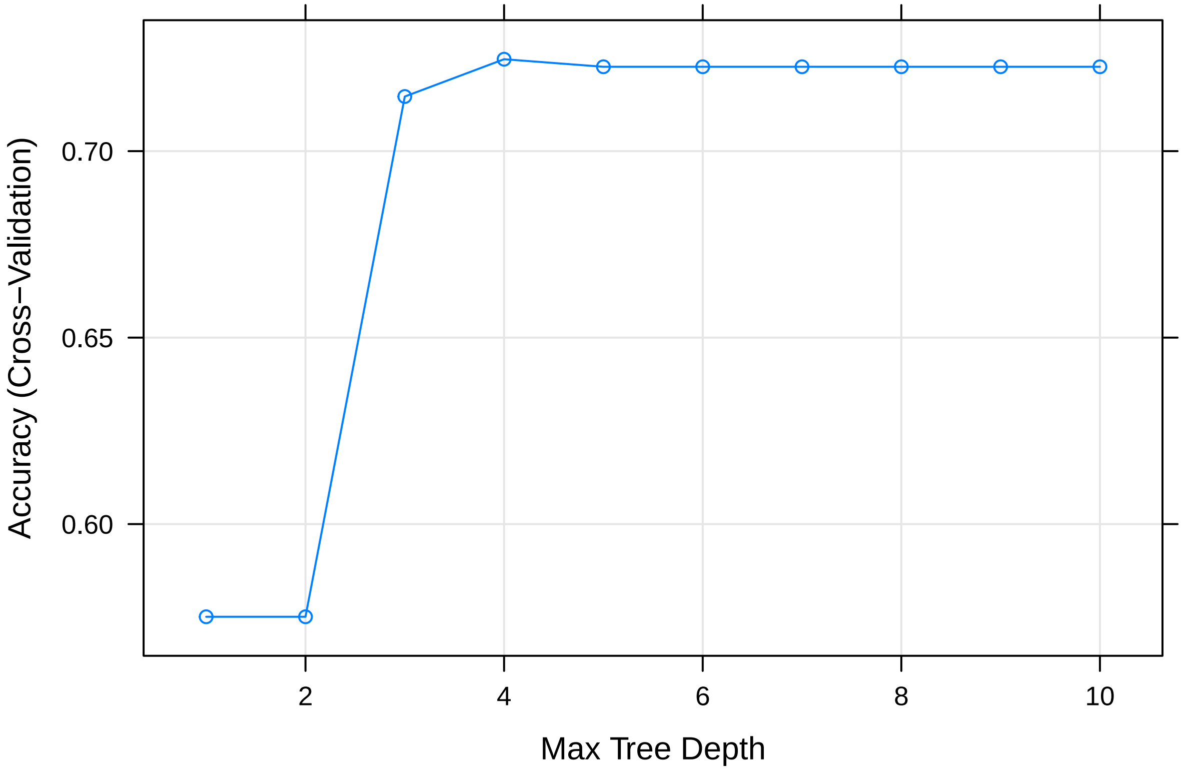
2. What is the overall classification error rate for the subtree? $18/105 = 0.1714286$

Note that if we cross validate the model (to find the best max depth), we get an accuracy of 0.724 (CV) and a maxdepth of 4. The *training* classification error rate will be much smaller (0.11) because the model is overfit to the training data. That is, above we didn't have a way to independently assess the accuracy of the model. Indeed, even in the cross validated case, we'd prefer to have an additional test dataset to assess the model accuracy.

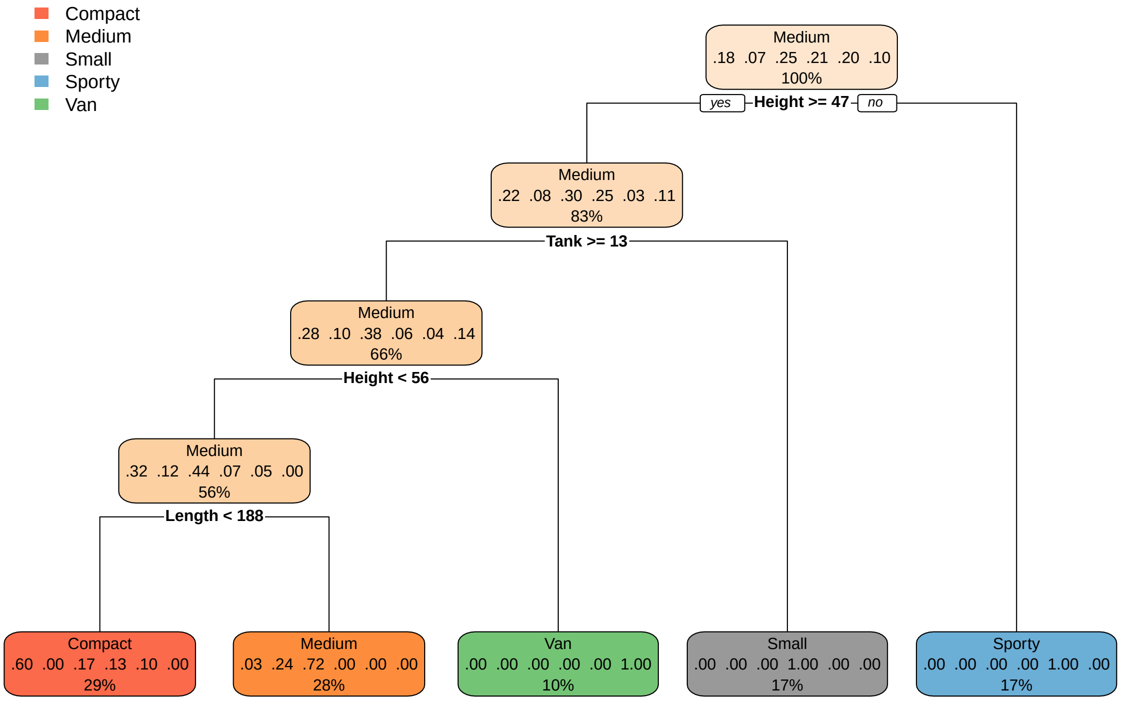
```r
set.seed(4747)
fitControl <- trainControl(method="cv")
cars.train<- train(Type ~ Length + Height + Price + Tank, data=car90,
                   method="rpart2", na.action = na.omit,
                   trControl = fitControl, tuneGrid= data.frame(maxdepth=1:10))
cars.train
```

```
## CART
##
## 105 samples
##   4 predictor
##   6 classes: 'Compact', 'Large', 'Medium', 'Small', 'Sporty', 'Van'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 96, 94, 94, 94, 94, 94, ...
## Resampling results across tuning parameters:
##
##   maxdepth  Accuracy   Kappa
##    1        0.5751515  0.4491877
##    2        0.5751515  0.4491877
##    3        0.7146465  0.6390385
##    4        0.7246465  0.6531411
##    5        0.7226263  0.6524134
##    6        0.7226263  0.6524134
##    7        0.7226263  0.6524134
##    8        0.7226263  0.6524134
##    9        0.7226263  0.6524134
##   10        0.7226263  0.6524134
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was maxdepth = 4.
```

```r
plot(cars.train)
```



```
rpart.plot(cars.train$finalModel)
```


Compact
Medium
Small
Sporty
Van
Medium
.18 .07 .25 .21 .20 .10
100%
yes
Height >= 47
no
Medium
.22 .08 .30 .25 .03 .11
83%
Tank >= 13
Medium
.28 .10 .38 .06 .04 .14
66%
Height < 56
Medium
.32 .12 .44 .07 .05 .00
56%
Length < 188
Compact
.60 .00 .17 .13 .10 .00
29%
Medium
.03 .24 .72 .00 .00 .00
28%
Van
.00 .00 .00 .00 .00 1.00
10%
Small
.00 .00 .00 1.00 .00 .00
17%
Sporty
.00 .00 .00 .00 1.00 .00
17%