R Zone 3: Chapter 10 and 11

Brandon Hufstetler, Garrett Alarcon, Jinan Andrews, Anson Cheng, Nick Forrest, and Nestor Herandez

15 July 2019

Chapter 10

CREATE THE DATA USING TABLE 10.3

```
new <- c(0.05,0.25)
A <- c(0.0467,  0.2471)
B <- c(0.0533,  0.1912)
C <- c(0.0917,  0.2794)
```

Create dataframe object and establish row/column names

```
data <- rbind(A, B, C)
dimnames(data) <- list(c("Dark", "Medium", "Light"),c("Age (MMN)", "Na/K (MMN)"))
```

Declare true classifications of A, B, and C.

```
trueclass <- c("Dark", "Medium", "Light")
```

Run KNN

```
# Requires package "class"
library(class)
knn <- knn(data, new, cl = trueclass, k = 3, prob = TRUE)
knn
```

```
## [1] Medium
## attr(,"prob")
## [1] 0.3333333
## Levels: Dark Light Medium
```

The predicted class is medium with a 33% probability.

Calculate the Euclidean distance

```
# Requires package "fields"
library(fields)

together <- rbind(new, data)
# The top row has the distances from New
rdist(together)
```

```
##             [,1]        [,2]       [,3]       [,4]
## [1,] 0.000000000 0.004393177 0.05889253 0.05102205
## [2,] 0.004393177 0.000000000 0.05628828 0.05539215
## [3,] 0.058892529 0.056288276 0.00000000 0.09619667
## [4,] 0.051022054 0.055392147 0.09619667 0.00000000
```

Stretch the axes

```
ds_newA <- sqrt((new[1]-A[1])^2 + (3*(new[2]-A[2]))^2)
ds_newB <- sqrt((new[1]-B[1])^2 + (3*(new[2]-B[2]))^2)
ds_newC <- sqrt((new[1]-C[1])^2 + (3*(new[2]-C[2]))^2)
```

Table 10.4

```
# Same thing as previous but adding BP as another variable
distance <- c(ds_newA, ds_newB, ds_newC)
BP <- c(120, 122, 130)
data <- cbind(BP, data, distance)
data
```

```
##         BP Age (MMN) Na/K (MMN)    distance
## Dark   120    0.0467     0.2471 0.009304837
## Medium 122    0.0533     0.1912 0.176430865
## Light  130    0.0917     0.2794 0.097560904
```

Locally Weighted Averaging

```
weights <- (1/(distance^2))
sum_wi <- sum(weights)
sum_wiyi <- sum(weights*data[,1])
yhat_new <- sum_wiyi/sum_wi
yhat_new
```

```
## [1] 120.0954
```

CLASSFIY RISK EXAMPLE: PREP THE DATA

```
#   Read in the ClassifyRisk dataset
setwd("~/IMGT680")
risk <- read.csv(file = "ClassifyRisk.txt", stringsAsFactors=FALSE, header=TRUE, sep=
"\t")

# Table <link   href="#urn:x-wiley:9783527333455:xml-component:c10:c10-tbl0005"/> contai
ns Records 51, 65, 79, 87, 124, 141, 150, 162, 163
```

Pull select samples from risk table along with selected variables

```
risk2 <- risk[c(51, 65, 79, 87, 124, 141, 150, 162), c(5, 1, 4, 6)]

# Categorical variables cannot be used in modeling
# Therefore turn marriage status into an indicator vairable
risk2$married.I <- ifelse(risk2$marital_status=="married", 1, 0)
risk2$single.I <- ifelse(risk2$marital_status=="single", 1, 0)

# Remove the two original categorical variables from the set
risk2 <- risk2[, -2];
risk2 <- risk2[, -2];

# Pull an observation out to be the test sameple and repeat the above process
new2    <- risk[163, c(5, 1, 4)]
new2$married.I <- 1;
new2$single.I <- 0;
new2 <- new2[, -2];
new2 <- new2[, -2];

# Establish response label
cll <- c(risk2[, 2])
```

ClassifyRisk example: KNN

```
# Train the KNN model and test it against the one sample pulled for the test set
knn2 <- knn(train = risk2[,c(1,3,4)], test = new2, cl = cll, k = 3)

#Display results
knn2
```

```
## [1] good risk
## Levels: bad loss good risk
```

The results from this KNN model predict the specific observation to be a good risk.

Chapter 11

Read in and prepare the data

```
setwd("~/IMGT680")
adult <- read.csv(file = "adult.txt", stringsAsFactors = TRUE)
```

Collapse some of the categories by giving them the same factor label

```
levels(adult$marital.status)
```

```
## [1] "Divorced"              "Married-AF-spouse"     "Married-civ-spouse"
## [4] "Married-spouse-absent" "Never-married"         "Separated"
## [7] "Widowed"
```

```
levels(adult$workclass)
```

```
## [1] "?"                "Federal-gov"      "Local-gov"
## [4] "Never-worked"     "Private"          "Self-emp-inc"
## [7] "Self-emp-not-inc" "State-gov"        "Without-pay"
```

```
levels(adult$marital.status)[2:4] <- "Married"
levels(adult$workclass)[c(2, 3, 8)] <- "Gov"
levels(adult$workclass)[c(5, 6)] <- "Self"
levels(adult$marital.status)
```

```
## [1] "Divorced"      "Married"       "Never-married" "Separated"
## [5] "Widowed"
```

```
levels(adult$workclass)
```

```
## [1] "?"            "Gov"          "Never-worked" "Private"
## [5] "Self"         "Without-pay"
```

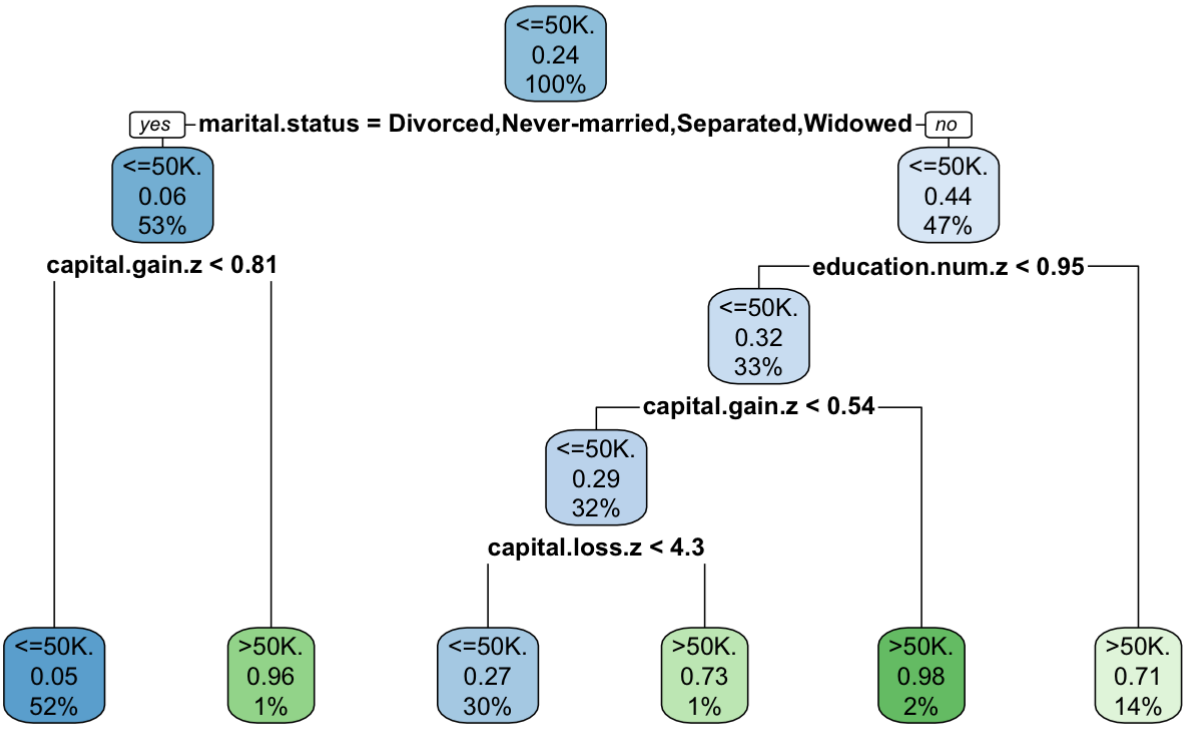
```
# Standardize the numeric variables
adult$age.z <- (adult$age - mean(adult$age))/sd(adult$age)
adult$education.num.z <- (adult$education.num - mean(adult$education.num))/sd(adult$educ
ation.num)
adult$capital.gain.z <- (adult$capital.gain - mean(adult$capital.gain))/sd(adult$capita
l.gain)
adult$capital.loss.z <- (adult$capital.loss - mean(adult$capital.loss))/sd(adult$capita
l.loss)
adult$hours.per.week.z <- (adult$hours.per.week - mean(adult$hours.per.week))/sd(adult$h
ours.per.week)

# Use predictors to classify whether or not a person's income is less than $50K
# Requires package “rpart”
library("rpart")
cartfit <- rpart(income ~ age.z + education.num.z + capital.gain.z + capital.loss.z + ho
urs.per.week.z + race + sex + workclass + marital.status, data = adult, method = "class"
)
print(cartfit)
```

```
## 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. (0.93
605751 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) *
```

Plot the decision tree

```
# Requires package “rpart.plot”
library("rpart.plot")
rpart.plot(cartfit, main = "Classification Tree")
```



The decision tree has 6 end nodes, and starts by splitting first on marital status. If yes to the split, then it only splits one more time with almost half the observations going to capital gain < 0.91. If no to marital status then three more splits are accomplished in order to classify the observations.

C5.0

```
# Requires package "C50"
library("C50")
names(adult)
```

```
##  [1] "age"            "workclass"        "demogweight"
##  [4] "education"      "education.num"    "marital.status"
##  [7] "occupation"     "relationship"     "race"
## [10] "sex"            "capital.gain"     "capital.loss"
## [13] "hours.per.week" "native.country"   "income"
## [16] "age.z"          "education.num.z"  "capital.gain.z"
## [19] "capital.loss.z" "hours.per.week.z"
```

```
x <- adult[,c(2, 6, 9, 10, 16, 17, 18, 19, 20)]
y <- adult$income
c50fit1 <- C5.0(x, y)
summary(c50fit1)
```

```
##
## Call:
## C5.0.default(x = x, y = y)
##
##
## C5.0 [Release 2.07 GPL Edition]      Sun Jul 14 11:47:38 2019
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 25000 cases (10 attributes) from undefined.data
##
## Decision tree:
##
## capital.gain.z > 0.7694287:
## :...hours.per.week.z > -0.4396555: >50K. (973/10)
## :   hours.per.week.z <= -0.4396555:
## :   :...age.z <= -0.8479775: <=50K. (6/1)
## :       age.z > -0.8479775: >50K. (86/3)
## capital.gain.z <= 0.7694287:
## :...marital.status in {Divorced,Never-married,Separated,Widowed}:
##     :...capital.loss.z > 5.282196:
##     :   :...capital.loss.z <= 5.646056: <=50K. (38/12)
##     :   :   capital.loss.z > 5.646056:
##     :   :   :...capital.loss.z <= 7.270963: >50K. (28)
##     :   :       capital.loss.z > 7.270963: <=50K. (8/1)
##     :   capital.loss.z <= 5.282196:
##     :   :...capital.gain.z > 0.4757047:
##     :       :...capital.gain.z <= 0.494004: >50K. (19)
##     :       :   capital.gain.z > 0.494004: <=50K. (67/6)
##     :       capital.gain.z <= 0.4757047:
##     :       :...education.num.z <= 0.7503064: <=50K. (10336/211)
##     :           education.num.z > 0.7503064:
##     :           :...education.num.z <= 1.532462:
##     :               :...hours.per.week.z <= 0.2107898: <=50K. (1598/109)
##     :               :   hours.per.week.z > 0.2107898:
##     :               :   :...sex = Female: <=50K. (314/49)
##     :               :       sex = Male:
##     :               :       :...education.num.z <= 1.141384: <=50K. (328/83)
##     :               :           education.num.z > 1.141384:
##     :               :           :...age.z > 1.197644: <=50K. (14/2)
##     :               :               age.z <= 1.197644:
##     :               :               :...age.z <= -0.4096299: <=50K. (15/1)
##     :               :                   age.z > -0.4096299: >50K. (50/14)
##     :               education.num.z > 1.532462:
##     :               :...age.z <= -0.4826879: <=50K. (57/7)
##     :                   age.z > -0.4826879:
##     :                   :...age.z > 1.124586: <=50K. (22/3)
##     :                       age.z <= 1.124586:
##     :                       :...marital.status in {Never-married,
##     :                           :                  Widowed}: >50K. (50/13)
##     :                           marital.status = Separated:
##     :                           :...sex = Female: <=50K. (6/2)
```

```
##     :                           :   sex = Male: >50K. (7/1)
##     :                           marital.status = Divorced:
##     :                           :...education.num.z <= 1.92354:
##     :                               :...sex = Female: <=50K. (8)
##     :                               :   sex = Male: >50K. (10/4)
##     :                               education.num.z > 1.92354:
##     :                               :...hours.per.week.z <= 1.755597: >50K. (9/2)
##     :                                   hours.per.week.z > 1.755597: <=50K. (2)
##     marital.status = Married:
##     :...capital.loss.z > 4.175664:
##         :...capital.loss.z <= 4.711484: >50K. (439/10)
##         :   capital.loss.z > 4.711484:
##         :   :...capital.loss.z <= 5.175032: <=50K. (48)
##         :       capital.loss.z > 5.175032:
##         :       :...education.num.z > 0.7503064: >50K. (47/1)
##         :           education.num.z <= 0.7503064:
##         :           :...education.num.z <= -0.8140052: <=50K. (7)
##         :               education.num.z > -0.8140052:
##         :               :...capital.loss.z > 5.803064: <=50K. (7)
##         :                   capital.loss.z <= 5.803064:
##         :                   :...capital.loss.z > 5.745743: >50K. (9)
##         :                       capital.loss.z <= 5.745743:
##         :                       :...age.z <= 0.613181: <=50K. (10)
##         :                           age.z > 0.613181: >50K. (13/4)
##         capital.loss.z <= 4.175664:
##         :...capital.gain.z > 0.5241912:
##             :...capital.gain.z <= 0.7118593: >50K. (77)
##             :   capital.gain.z > 0.7118593:
##             :   :...capital.gain.z <= 0.7246822: >50K. (4)
##             :       capital.gain.z > 0.7246822: <=50K. (4)
##             capital.gain.z <= 0.5241912:
##             :...education.num.z > 0.7503064:
##                 :...capital.loss.z > 1.342044: <=50K. (25/5)
##                 :   capital.loss.z <= 1.342044:
##                 :   :...capital.gain.z > 0.2690694:
##                 :       :...capital.gain.z <= 0.4023739: <=50K. (25)
##                 :       :   capital.gain.z > 0.4023739:
##                 :       :   :...capital.gain.z <= 0.4444489: >50K. (12/1)
##                 :       :       capital.gain.z > 0.4444489: <=50K. (17)
##                 :       capital.gain.z <= 0.2690694:
##                 :       :...hours.per.week.z <= -0.6835725:
##                 :           :...sex = Female: >50K. (69/28)
##                 :           :   sex = Male:
##                 :           :   :...education.num.z <= 1.532462: <=50K. (140/31)
##                 :           :       education.num.z > 1.532462: >50K. (31/12)
##                 :           hours.per.week.z > -0.6835725:
##                 :           :...age.z <= -0.4096299:
##                 :               :...age.z <= -0.9940934: <=50K. (50/14)
##                 :               :   age.z > -0.9940934: [S1]
##                 :               age.z > -0.4096299:
##                 :               :...race in {Black,White}: >50K. (1747/507)
##                 :                   race = Other: <=50K. (6/1)
##                 :                   race = Amer-Indian-Eskimo:
##                 :                   :...hours.per.week.z <= 0.2920955: <=50K. (4)
```

```
##                 :                    :   hours.per.week.z > 0.2920955: >50K. (3)
##                 :                    race = Asian-Pac-Islander:
##                 :                    :...sex = Male: >50K. (95/41)
##                 :                        sex = Female:
##                 :                        :...age.z <= 0.1748335: >50K. (6/1)
##                 :                            age.z > 0.1748335: <=50K. (15/2)
##                 education.num.z <= 0.7503064:
##                 :...capital.loss.z > 3.68221: <=50K. (92)
##                     capital.loss.z <= 3.68221:
##                     :...capital.gain.z > 0.4444489: <=50K. (47)
##                         capital.gain.z <= 0.4444489:
##                         :...capital.gain.z > 0.4023739:
##                             :...hours.per.week.z <= 0.9425408: >50K. (41/5)
##                             :   hours.per.week.z > 0.9425408: <=50K. (4/1)
##                             capital.gain.z <= 0.4023739:
##                             :...capital.gain.z > 0.2690694: <=50K. (131)
##                                 capital.gain.z <= 0.2690694:
##                                 :...capital.gain.z > 0.2543766: >50K. (50/2)
##                                     capital.gain.z <= 0.2543766:
##                                     :...capital.gain.z > -0.09184103: <=50K. (118)
##                                         capital.gain.z <= -0.09184103:
##                                         :...hours.per.week.z <= -0.5209612: <=50K. (7
36/92)
##                                             hours.per.week.z > -0.5209612:
##                                             :...age.z <= -0.2635141: <=50K. (2178/40
8)
##                                                 age.z > -0.2635141: [S2]
##
## SubTree [S1]
##
## race in {Asian-Pac-Islander,Other}: <=50K. (26/10)
## race in {Amer-Indian-Eskimo,White}: >50K. (377/162)
## race = Black:
## :...education.num.z <= 1.141384: >50K. (19/8)
##     education.num.z > 1.141384: <=50K. (2)
##
## SubTree [S2]
##
## race in {Amer-Indian-Eskimo,Other}: <=50K. (55/7)
## race in {Asian-Pac-Islander,Black,White}:
## :...education.num.z <= -0.4229273: <=50K. (2659/777)
##     education.num.z > -0.4229273:
##     :...hours.per.week.z > 0.2107898:
##         :...workclass in {?,Gov,Never-worked,Private}: >50K. (410.8/181.4)
##         :   workclass = Without-pay: <=50K. (1)
##         :   workclass = Self:
##         :   :...education.num.z <= 0.3592285: <=50K. (175.2/74.5)
##         :       education.num.z > 0.3592285:
##         :       :...hours.per.week.z <= 1.349069: >50K. (11/2)
##         :           hours.per.week.z > 1.349069: <=50K. (4)
##         hours.per.week.z <= 0.2107898:
##         :...workclass in {?,Gov,Never-worked,Without-pay}: <=50K. (190/93.5)
##             workclass = Self:
##             :...sex = Female: >50K. (8.4/2.3)
```

```
##             :   sex = Male: <=50K. (85.6/33.6)
##             workclass = Private:
##             :...age.z > 0.613181:
##                 :...age.z <= 1.70905: >50K. (206.5/94.8)
##                 :   age.z > 1.70905: <=50K. (20.7/3.7)
##                 age.z <= 0.613181:
##                 :...sex = Male: <=50K. (337.4/118.7)
##                     sex = Female:
##                     :...education.num.z <= -0.03184938: <=50K. (41.7/17)
##                         education.num.z > -0.03184938: >50K. (11.7/2)
##
##
## Evaluation on training data (25000 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      78 3286(13.1%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##   17902  1114    (a): class <=50K.
##    2172  3812    (b): class >50K.
##
##
##  Attribute usage:
##
##  100.00% capital.gain.z
##   95.74% marital.status
##   95.74% capital.loss.z
##   92.81% education.num.z
##   52.65% hours.per.week.z
##   36.44% age.z
##   26.07% race
##    6.43% sex
##    5.90% workclass
##
##
## Time: 0.1 secs
```

The first thing to notice is the error percent, which is 13.1 %. Of the 9 variables used, only 4 seemed to be of high importance. These were capital gain, marital status, capital loss, and education num. All in order of decreasing importance, but all four variables have higher than 90% importance. The fifth highest is hours per week, which had a importance score of 52.65 %.

C5.0 - Pruned

```
c50fit2 <- C5.0(x, y, control = C5.0Control(CF=.1))
summary(c50fit2)
```

```
##
## Call:
## C5.0.default(x = x, y = y, control = C5.0Control(CF = 0.1))
##
##
## C5.0 [Release 2.07 GPL Edition]      Sun Jul 14 11:47:39 2019
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 25000 cases (10 attributes) from undefined.data
##
## Decision tree:
##
## capital.gain.z > 0.7694287: >50K. (1065/18)
## capital.gain.z <= 0.7694287:
## :...marital.status in {Divorced,Never-married,Separated,Widowed}:
##     :...capital.loss.z > 5.282196:
##     :   :...capital.loss.z <= 5.646056: <=50K. (38/12)
##     :   :   capital.loss.z > 5.646056: >50K. (36/7)
##     :   capital.loss.z <= 5.282196:
##     :   :...capital.gain.z > 0.4757047:
##     :       :...capital.gain.z <= 0.494004: >50K. (19)
##     :       :   capital.gain.z > 0.494004: <=50K. (67/6)
##     :       capital.gain.z <= 0.4757047:
##     :       :...education.num.z <= 0.7503064: <=50K. (10336/211)
##     :           education.num.z > 0.7503064:
##     :           :...education.num.z <= 1.532462: <=50K. (2319/280)
##     :               education.num.z > 1.532462:
##     :               :...age.z <= -0.4826879: <=50K. (57/7)
##     :                   age.z > -0.4826879:
##     :                   :...age.z <= 1.124586: >50K. (92/34)
##     :                       age.z > 1.124586: <=50K. (22/3)
##     marital.status = Married:
##     :...capital.loss.z > 4.175664:
##         :...capital.loss.z <= 4.711484: >50K. (439/10)
##         :   capital.loss.z > 4.711484:
##         :   :...capital.loss.z <= 5.175032: <=50K. (48)
##         :       capital.loss.z > 5.175032:
##         :       :...education.num.z <= 0.7503064: <=50K. (46/18)
##         :           education.num.z > 0.7503064: >50K. (47/1)
##         capital.loss.z <= 4.175664:
##         :...capital.gain.z > 0.5241912: >50K. (85/4)
##             capital.gain.z <= 0.5241912:
##             :...education.num.z <= 0.7503064:
##                 :...capital.gain.z > 0.4444489: <=50K. (47)
##                 :   capital.gain.z <= 0.4444489:
##                 :   :...capital.gain.z > 0.4023739: >50K. (45/8)
##                 :       capital.gain.z <= 0.4023739:
##                 :       :...capital.gain.z > 0.2690694: <=50K. (131)
##                 :           capital.gain.z <= 0.2690694:
##                 :           :...capital.gain.z <= 0.2543766: <=50K. (7342/1991)
##                 :               capital.gain.z > 0.2543766: >50K. (50/2)
```

```
##                 education.num.z > 0.7503064:
##                 :...capital.loss.z > 1.342044: <=50K. (25/5)
##                     capital.loss.z <= 1.342044:
##                     :...capital.gain.z > 0.2690694:
##                         :...capital.gain.z <= 0.4023739: <=50K. (25)
##                         :   capital.gain.z > 0.4023739:
##                         :   :...capital.gain.z <= 0.4444489: >50K. (12/1)
##                         :       capital.gain.z > 0.4444489: <=50K. (17)
##                         capital.gain.z <= 0.2690694:
##                         :...hours.per.week.z <= -0.6835725:
##                             :...sex = Female: >50K. (69/28)
##                             :   sex = Male:
##                             :   :...education.num.z <= 1.532462: <=50K. (140/31)
##                             :       education.num.z > 1.532462: >50K. (31/12)
##                             hours.per.week.z > -0.6835725:
##                             :...age.z > -0.4096299: >50K. (1876/571)
##                                 age.z <= -0.4096299:
##                                 :...age.z <= -0.9940934: <=50K. (50/14)
##                                     age.z > -0.9940934: >50K. (424/188)
##
##
## Evaluation on training data (25000 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      30 3462(13.8%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##   18132   884    (a): class <=50K.
##    2578  3406    (b): class >50K.
##
##
##  Attribute usage:
##
##  100.00% capital.gain.z
##   95.74% marital.status
##   95.74% capital.loss.z
##   92.81% education.num.z
##   10.36% hours.per.week.z
##   10.08% age.z
##    0.96% sex
##
##
## Time: 0.1 secs
```

While the error rate increased to 13.8 % after pruning some of the tree, it is not a large increase in the error rate. Especially considering the speed of the calculation dropped from 0.3 to 0.2 seconds. While still increadibly fast, it still represents a 33% reduction in computation time. As far as important variables, nothing drastic changed in the

top four reported previously, and they are still above 90%. What's interesting is the fifth most important variable, hours per week, dropped to 10.36% which is a drastic redution from the previous model.