Decision Tree CART and C5.0 Model

Amin Fesharaki

11/7/2021

For Exercises 14–20, work with the adult_ch6_training and adult_ch6_test data sets. Use either Python or R to solve each problem.

```
#Import Datasets
adult_test <- read.csv(file = "/Users/datascience/Desktop/ADS 502 Data Sets/Website Data Sets/adult_ch6
adult_train <- read.csv(file = "/Users/datascience/Desktop/ADS 502 Data Sets/Website Data Sets/adult_ch
library(C50)
library(rpart); library(rpart.plot)
```

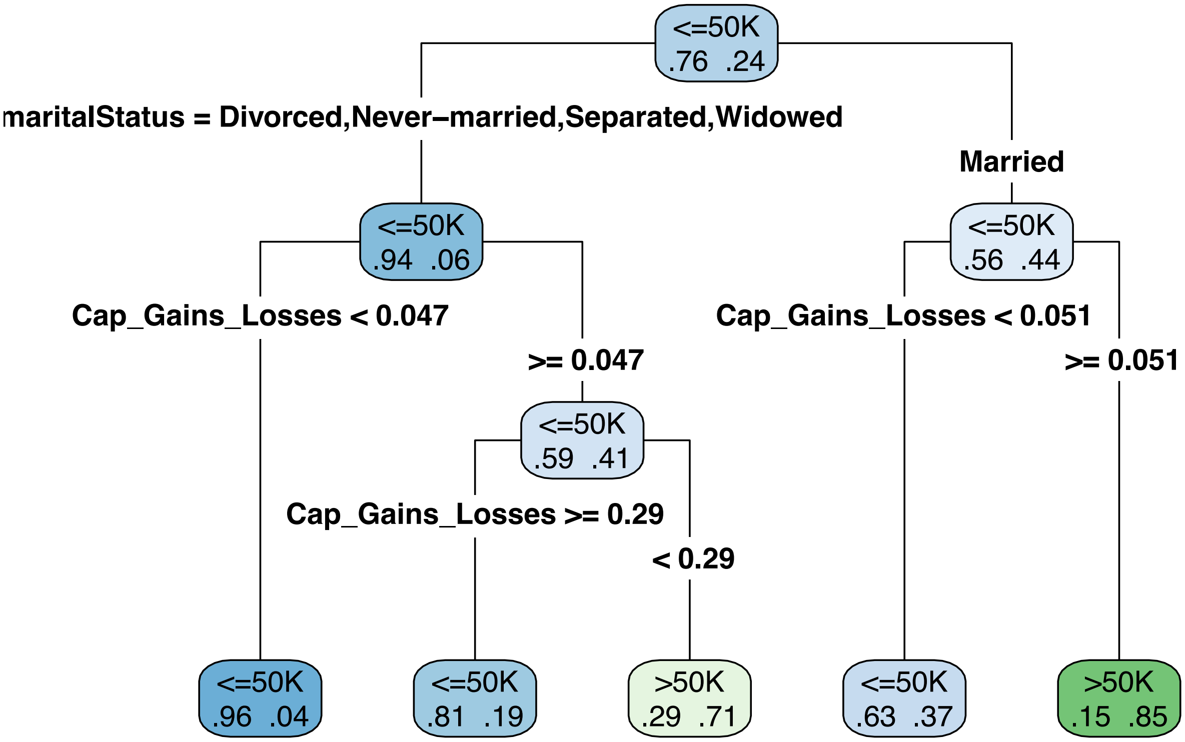
14. Create a CART model using the training data set that predicts income using marital status and capital gains and losses. Visualize the decision tree (that is, provide the decision tree output). Describe the first few splits in the decision tree.

```
# Change the name to eliminate spaces
colnames(adult_train)[1] <- "maritalStatus"

# Change categorical variables into factors
adult_train$Income <- factor(adult_train$Income)
adult_train$maritalStatus <- factor(adult_train$maritalStatus)

# Build CART model
cart  <- rpart(formula = Income ~ maritalStatus + Cap_Gains_Losses, data = adult_train, method = "class

# Plot Cart Model
rpart.plot(cart, type = 4, extra = 4)
```



```
# Predictions
X = data.frame(maritalStatus = adult_train$maritalStatus, Cap_Gains_Losses = adult_train$Cap_Gains_Losse
Pred = predict(object = cart, newdata = X, type = "class")
```

** The root node tells us that there are 24% of the records with high income (>50K) and 76% with low income (<50k). Then the node splits off depending if the individual is married or not. For example, the non married node (left side of root node) shows that 6% has high income and the married side (right side) shows that 44% have high income. Then an additional split happens depending on your capital gains and losses. **

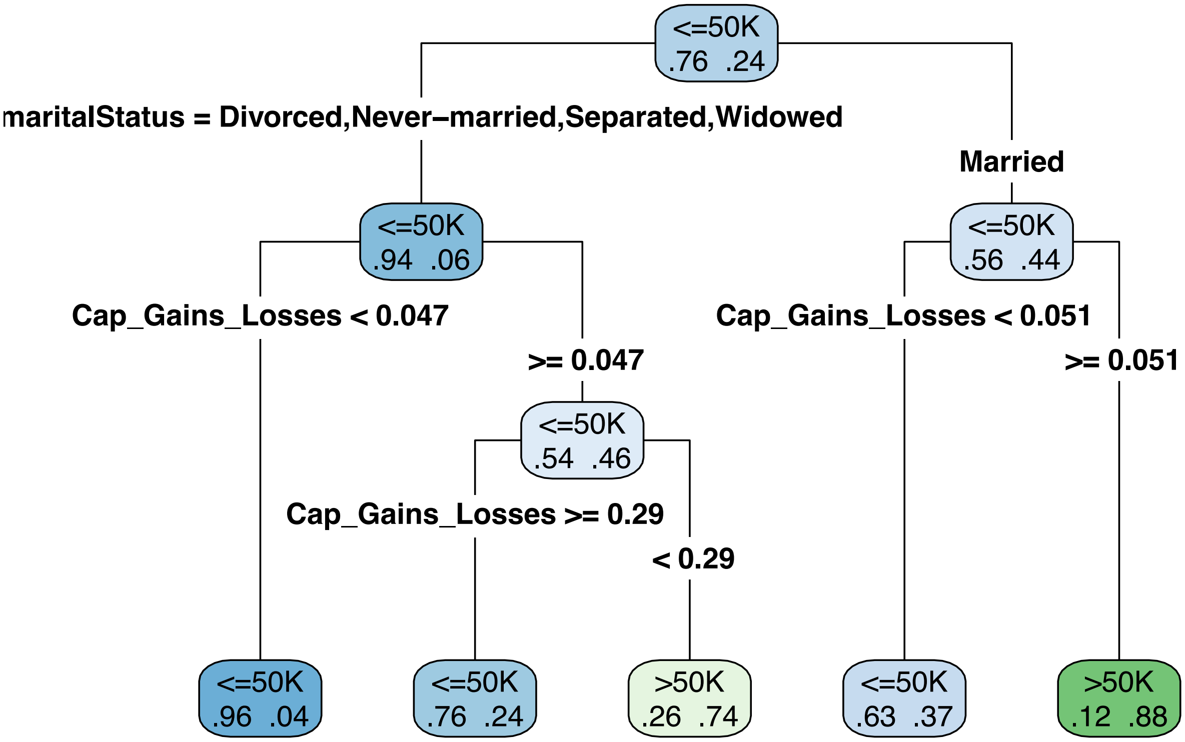
15. Develop a CART model using the test data set that utilizes the same target and predictor variables. Visualize the decision tree. Compare the decision trees. Does the test data result match the training data result?

```
# Change the name to eliminate spaces
colnames(adult_test)[1] <- "maritalStatus"

# Change categorical variables into factors
adult_test$Income <- factor(adult_test$Income)
adult_test$maritalStatus <- factor(adult_test$maritalStatus)

# Build CART model
cart_t  <- rpart(formula = Income ~ maritalStatus + Cap_Gains_Losses, data = adult_test, method = "clas

# Plot Cart Model
rpart.plot(cart_t, type = 4, extra = 4)
```



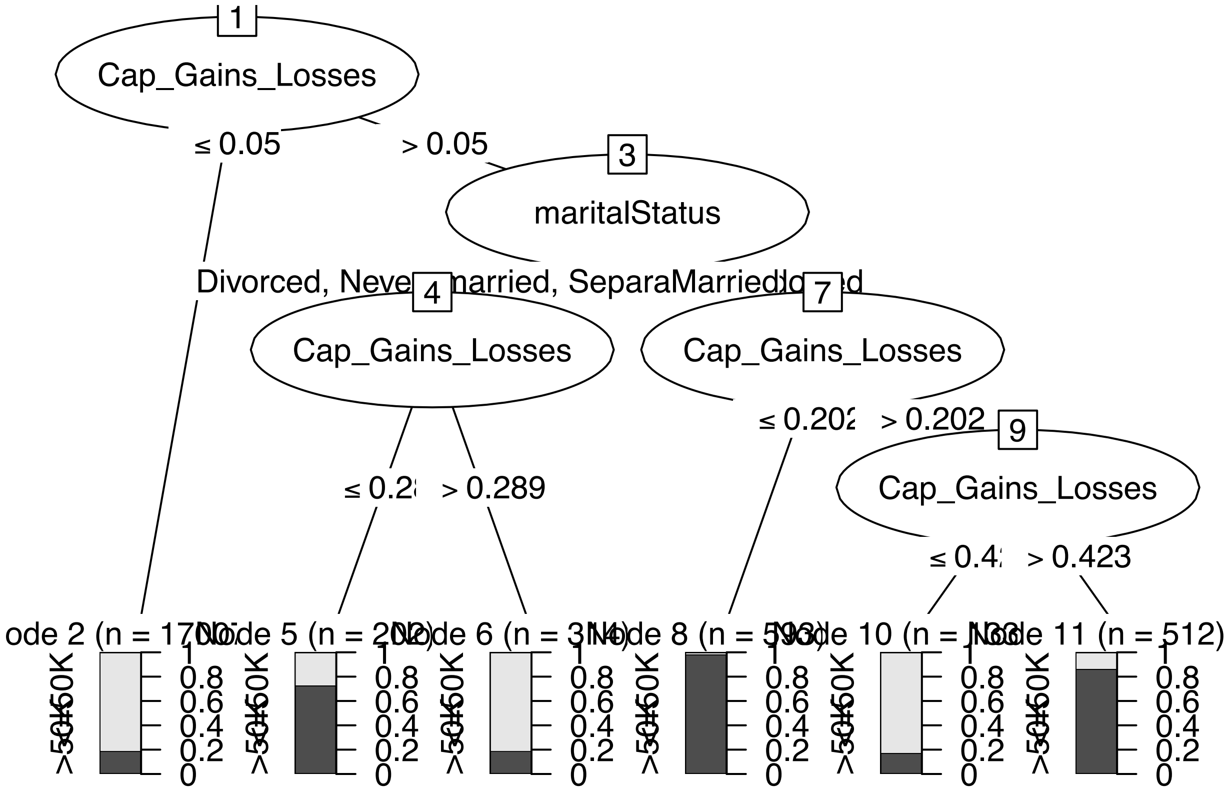
```
# Predictions
X_t = data.frame(maritalStatus = adult_test$maritalStatus, Cap_Gains_Losses = adult_test$Cap_Gains_Losse
Pred_t = predict(object = cart_t, newdata = X_t, type = "class")
```

- The root and internal node are the same between the training and testing data. However, there are some differences in values for the leaf nodes. *

16. Use the training data set to build a C5.0 model to predict income using marital status and capital gains and losses. Specify a minimum of 75 cases per terminal node. Visualize the decision tree. Describe the first few splits in the decision tree.

```
# C5 model
c5 <- C5.0(formula = Income ~ maritalStatus + Cap_Gains_Losses, data = adult_train, control = C5.0Contro

# Graph
plot(c5)
```



```
# Predict
Predictions2 = predict(object = c5, newdata = X)
```

** The root node for the C5.0 model is the Capital Gain losses. From there it will either terminate (<.05) or move onto the next node which is the marital status. Then it will split depending if you are married or not. Then the node will split again depending on the Capital Gain and Losses.**

17. How does your C5.0 model compare to the CART model? Describe the similarities and differences.

** Both models split the data according to its criteria, however there are many differences between the two. One difference is that there are 11 nodes in the C5.0 model while only 9 nodes in the CART model. Moreover, the C5.0 model immediately sends 90.7% of the data into the leaf node 2. The rest of the nodes work with 9.3% of the data. On the other hand, the Cart node split based on Marital status instead of Cap Gains Losses resulting in a more balance split. The splitting nature is mostly caused by the Gini index rather than C5.0's entropy reduction.