Decision Tree Model

Anushree Shivarudrappa

June 15, 2016

1. Pre-Processing

```
library(data.table)
library(ggplot2)
library(dplyr)
library(scales)
library(RColorBrewer)
library(tidyr)
library(caTools)
library(rpart)
library(rpart.plot)
library(ROCR)
library(randomForest)
library(tree)
library(caret)
library(e1071)
```

2. Data Loading

```
Death_US <- fread("DeathRecords.csv", header = T)
```

3. Selecting dataset for model

```
# separates natural death
Death_US_natural <- Death_US[Death_US$MannerOfDeath == 7, ]
```

Select required variables

```
require(MASS)
require(dplyr)
natural_sub <- Death_US_natural %>% dplyr::select(Education2003Revision, Sex, Age,
                      InfantAgeRecode22,
                      PlaceOfDeathAndDecedentsStatus, MaritalStatus, InjuryAtWork,
                      Autopsy, ActivityCode, PlaceOfInjury, Icd10Code,CauseRecode358,
                      CauseRecode113, InfantCauseRecode130,CauseRecode39,
                      NumberOfEntityAxisConditions,NumberOfRecordAxisConditions,Race)

str(natural_sub)
```

```
## Classes 'data.table' and 'data.frame':   2059933 obs. of  18 variables:
##  $ Education2003Revision         : int  2 2 7 6 3 5 4 4 3 3 ...
##  $ Sex                           : chr  "M" "M" "F" "M" ...
##  $ Age                           : int  87 58 75 74 64 93 82 55 86 79 ...
##  $ InfantAgeRecode22             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PlaceOfDeathAndDecedentsStatus: int  4 4 4 6 4 6 4 7 4 6 ...
##  $ MaritalStatus                 : chr  "M" "D" "W" "D" ...
##  $ InjuryAtWork                  : chr  "U" "U" "U" "U" ...
##  $ Autopsy                       : chr  "N" "N" "N" "N" ...
##  $ ActivityCode                  : int  99 99 99 99 99 99 99 99 99 99 ...
##  $ PlaceOfInjury                 : int  99 99 99 99 99 99 99 99 99 99 ...
##  $ Icd10Code                     : chr  "I64" "I250" "J449" "I48" ...
##  $ CauseRecode358                : int  238 214 267 228 214 280 215 214 175 225 ...
##  $ CauseRecode113                : int  70 62 86 68 62 111 63 62 111 68 ...
##  $ InfantCauseRecode130          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CauseRecode39                 : int  24 21 28 22 21 37 21 21 37 22 ...
##  $ NumberOfEntityAxisConditions  : int  1 3 2 3 1 5 4 2 1 2 ...
##  $ NumberOfRecordAxisConditions  : int  1 3 2 3 1 5 3 2 1 2 ...
##  $ Race                          : int  1 1 1 1 1 1 1 2 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Converting Character variable into Integer variable

```
natural_sub$Sex <- as.integer(as.factor(natural_sub$Sex))
natural_sub$MaritalStatus <- as.integer(as.factor(natural_sub$MaritalStatus))
natural_sub$InjuryAtWork <- as.integer(as.factor(natural_sub$InjuryAtWork))
natural_sub$Autopsy <- gsub("n", "N", natural_sub$Autopsy)
natural_sub$Autopsy <- as.integer(as.factor(natural_sub$Autopsy))
natural_sub$Icd10Code <- as.integer(as.factor(natural_sub$Icd10Code))
```

As we analyzed, the feature variables are "Age + InfantAgeRecode22 + PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode + PlaceOfInjury + NumberOfRecordAxisConditions + NumberOfEntityAxis-Conditions"

```
# Since the decision tree support till 32 levels removing 7 levels which has less entries
table(factor(natural_sub$CauseRecode39))
```

```
##
##      1      2      3      5      6      7      8      9     10     11
##    366     37   5619   9053  43839  33847 133412  34621  23359  23422
##     12     13     14     15     16     17     20     21     22     23
##  25734  17116  19671 133276  63721  75552  37415 310848 175752  23704
##     24     25     26     27     28     29     30     31     32     33
## 111664   5426  16551  45801 125752   2519  31595  41369   1000   9930
##     34     35     36     37     38     39     40     41     42
##   8110    414  23035 433081    212  13088      8      5      9
```

```
CauseExtraRemove <- natural_sub[, natural_sub$CauseRecode39 %in% c(2, 40, 41, 42, 38, 35, 1)]
table(CauseExtraRemove)
```

```
## CauseExtraRemove
```

```
##   FALSE    TRUE
## 2058882    1051
```

```
# remove the 7 factors levels from Death_US_natural dataset
natural_sub <- natural_sub[!(CauseExtraRemove)]
nrow(natural_sub)
```

```
## [1] 2058882
```

```
# model data
modeldata <- natural_sub

# We will do a random 70:30 split in our data set (70% will be for training models,
# 30% to evaluate them)
set.seed(111)
# randomly pick 70% of the number of observations
index <- sample.split(modeldata$CauseRecode39, SplitRatio = 0.7)
# subset data to include only the elements in the index
train <- subset(modeldata, index==T)
nrow(train)
```

```
## [1] 1441215
```

```
# subset data to include all but the elements in the index
test <- subset(modeldata, index==F)
nrow(test)
```

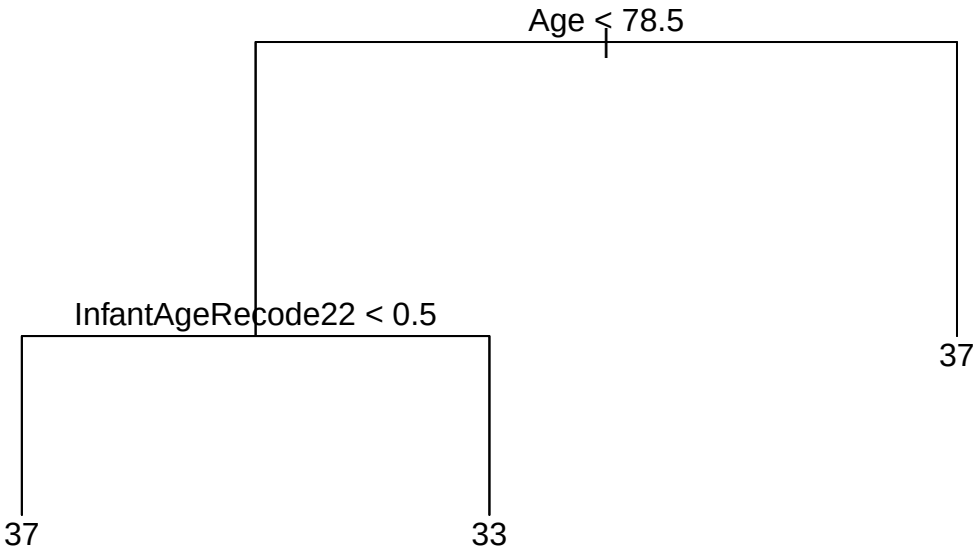
```
## [1] 617667
```

```
# take a copy of ICD10Code of test set and remove the variable from test set
Cause39 <- test$CauseRecode39
test$CauseRecode39 <- NULL
```

Model Decision Tree

```
model_tree <- tree(as.factor(CauseRecode39) ~ Age + InfantAgeRecode22 +
                     PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                      PlaceOfInjury + NumberOfRecordAxisConditions +
                      NumberOfEntityAxisConditions, train)

# plot model_tree
plot(model_tree)
text(model_tree, pretty = 0)
```




```
# Predict the test dataset using model
predict_ICD2 <- predict(model_tree, newdata = test, type = "class")
# confusion matrix
conf_matrix2 <- table(predict_ICD2, Cause39)
```

Model Accuracy

```
sum(diag(conf_matrix2)) / nrow(test)
```

```
## [1] 0.2143631
```

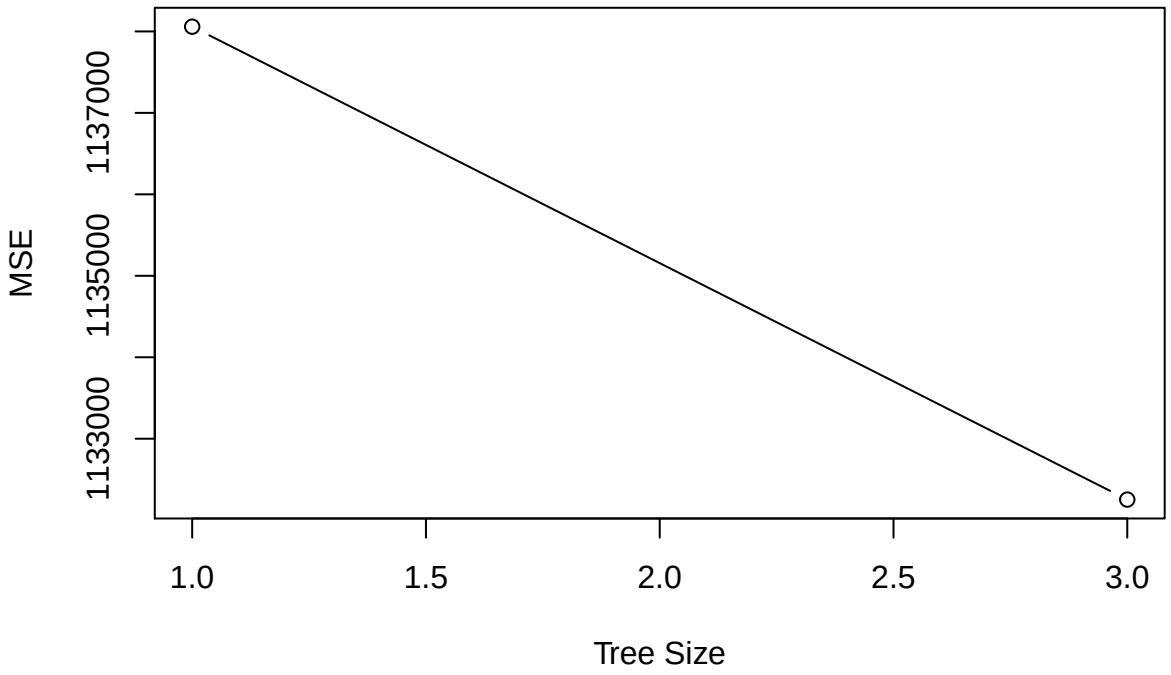
Prune tree

Prune back the tree to avoid overfitting the data. Typically, you will want to select a tree size that minimizes the cross-validated error

```
# cross validation to check where to stop prunning
cvtree <- cv.tree(model_tree, FUN = prune.misclass)
names(cvtree)
```

```
## [1] "size"   "dev"    "k"      "method"
```

```
plot(cvtree$size, cvtree$dev, type = "b",
     xlab = "Tree Size",
     ylab = "MSE")
```



Since the lowest deviation is at tree size 3 which we already have in our model, there is no need to prune the tree