Random Forest 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,
                      MannerOfDeath,
                      Autopsy, ActivityCode, PlaceOfInjury, Icd10Code,CauseRecode358,
                      CauseRecode113, InfantCauseRecode130,CauseRecode39,
                      NumberOfEntityAxisConditions,NumberOfRecordAxisConditions,Race)
```

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 + NumberOfEntityAxisConditions"

```
# 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 2 :: Random Forest

```
model_forest1 <- randomForest(as.factor(CauseRecode39) ~ Age + InfantAgeRecode22 +
                      PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                        PlaceOfInjury + NumberOfRecordAxisConditions +
                        NumberOfEntityAxisConditions,
                      data = train[1:600000, ], nodesize = 25, ntree = 1501)

model_forest2 <- randomForest(as.factor(CauseRecode39) ~ Age + InfantAgeRecode22 +
                      PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                        PlaceOfInjury + NumberOfRecordAxisConditions +
                        NumberOfEntityAxisConditions,
                      data = train[600001:1200000, ],
                      nodesize = 25, ntree = 1501)

model_forest <- combine(model_forest1,model_forest2)

# Predict the test dataset using random forest model
predict_forest <- predict(model_forest, newdata = test)
# confusion matrix
conf_matrix <- table(predict_forest, Cause39)
```

Model Accuracy

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

```
## [1] 0.2590312
```