Feature Selection in CKME136

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

#

# Load required packages

require(plyr)

## Loading required package: plyr

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(FSelector)  
library(FSelectorRcpp)  
library(ElemStatLearn)

##   
## Attaching package: 'ElemStatLearn'

## The following object is masked from 'package:plyr':  
##   
## ozone

#library(corrplot)

# Load data

fcpp <- read.csv("C:/Users/YENN/Desktop/UST/FARS2016N/accident2016.csv", header = T, stringsAsFactors = F)

# Merge YEAR, MONTH, DAY, HOUR, MINUTE into Timestamps

fcpp$TIMESTAMP <- with(fcpp, ISOdatetime(YEAR, MONTH, DAY, HOUR, MINUTE, sec = 0, tz=""))

# Remove TWAY\_ID2 attribute, the only variable with missing values: <sum(is.na (accs$TWAY\_ID2))> and TWAY\_ID, not appropiate for the research project

# Remove YEAR, MONTH, DAY, HOUR, MINUTE attributes - it’s been merged into Timestamps 12:14

# Remove WEATHER1, WEATHER2 attributes, are duplicate of the original WEATHER

# Remove RAIL attribute, no relevant to the research

## accs2016 <- accs[,-c(1:2,10:11,12:14,16:17,23:24,37:38,41)]

fcpp16 <- fcpp[,-c(1:2,10:14,16:17,23:24,37:38,41,53)]

# Values of the FATALS attributes

table(fcpp16$FATALS)

##   
## 1 2 3 4 5 6 9   
## 31984 2033 315 80 19 7 1

fcpp16$FATALS <- mapvalues(fcpp16$FATALS, from = c("1", "2", "3", "4", "5", "6", "9"), to = c(0,1,1,1,1,1,1))  
  
table(fcpp16$FATALS)

##   
## 0 1   
## 31984 2455

The above result shows imbalanced data.

fcpp16$FATALS <- as.factor(fcpp16$FATALS)

##### Using FSelectorRcpp

# Using all the variables before pre-process using FSelectorRcpp

#library(FSelectorRcpp)  
x <- fcpp  
y <- fcpp$FATALS  
information\_gain(x=x,y=y)

## Warning in .information\_gain.data.frame(x, y, type, equal, threads): There  
## are missing values in your data. information\_gain will remove them.

## Warning in .information\_gain.data.frame(x, y, type, equal, threads):  
## Dependent variable is a numeric! It will be converted to factor with simple  
## factor(y). We do not discretize dependent variable in FSelectorRcpp by  
## default! You can choose equal frequency binning discretization by setting  
## equal argument to TRUE.

## attributes importance  
## 1 STATE 3.857525e-03  
## 2 ST\_CASE 2.996407e-01  
## 3 VE\_TOTAL 1.069172e-02  
## 4 VE\_FORMS 1.069352e-02  
## 5 PVH\_INVL 3.158302e-04  
## 6 PEDS 1.563234e-02  
## 7 PERNOTMVIT 1.446612e-02  
## 8 PERMVIT 4.942634e-02  
## 9 PERSONS 5.009079e-02  
## 10 COUNTY 1.259545e-02  
## 11 CITY 3.557081e-02  
## 12 DAY 2.402551e-03  
## 13 MONTH 9.622124e-04  
## 14 YEAR 0.000000e+00  
## 15 DAY\_WEEK 1.364694e-03  
## 16 HOUR 2.533324e-03  
## 17 MINUTE 3.301577e-03  
## 18 NHS 1.280612e-03  
## 19 RUR\_URB 2.448515e-03  
## 20 FUNC\_SYS 1.954804e-03  
## 21 RD\_OWNER 2.989516e-03  
## 22 ROUTE 3.208865e-03  
## 23 TWAY\_ID 1.551950e-01  
## 24 TWAY\_ID2 6.754256e-02  
## 25 MILEPT 4.425825e-02  
## 26 LATITUDE 0.000000e+00  
## 27 LONGITUD 0.000000e+00  
## 28 SP\_JUR 7.369216e-04  
## 29 HARM\_EV 1.722843e-02  
## 30 MAN\_COLL 1.335743e-02  
## 31 RELJCT1 3.346646e-05  
## 32 RELJCT2 1.878310e-03  
## 33 TYP\_INT 8.921803e-04  
## 34 WRK\_ZONE 3.513458e-04  
## 35 REL\_ROAD 1.811018e-03  
## 36 LGT\_COND 8.792855e-04  
## 37 WEATHER1 7.179412e-04  
## 38 WEATHER2 3.610412e-04  
## 39 WEATHER 7.077501e-04  
## 40 SCH\_BUS 9.525406e-05  
## 41 RAIL 1.915302e-03  
## 42 NOT\_HOUR 2.383521e-03  
## 43 NOT\_MIN 4.276626e-03  
## 44 ARR\_HOUR 2.404990e-03  
## 45 ARR\_MIN 4.008167e-03  
## 46 HOSP\_HR 3.324157e-03  
## 47 HOSP\_MN 5.423978e-03  
## 48 CF1 1.356945e-03  
## 49 CF2 3.771647e-04  
## 50 CF3 3.233057e-04  
## 51 FATALS 2.996407e-01  
## 52 DRUNK\_DR 2.338857e-03  
## 53 TIMESTAMP 0.000000e+00

# Data has missing values…

#library(FSelectorRcpp)  
x2 <- fcpp16  
y2 <- fcpp16$FATALS  
information\_gain(x=x2,y=y2)

## attributes importance  
## 1 VE\_TOTAL 9.838390e-03  
## 2 VE\_FORMS 9.937294e-03  
## 3 PVH\_INVL 8.014738e-05  
## 4 PEDS 1.492545e-02  
## 5 PERNOTMVIT 1.385067e-02  
## 6 PERMVIT 4.235427e-02  
## 7 PERSONS 4.302656e-02  
## 8 DAY\_WEEK 8.047761e-04  
## 9 NHS 9.354703e-04  
## 10 RUR\_URB 2.080219e-03  
## 11 FUNC\_SYS 1.307897e-03  
## 12 RD\_OWNER 2.255107e-03  
## 13 ROUTE 2.391394e-03  
## 14 MILEPT 2.852640e-02  
## 15 LATITUDE 0.000000e+00  
## 16 LONGITUD 4.274073e-04  
## 17 SP\_JUR 5.015114e-04  
## 18 HARM\_EV 1.506612e-02  
## 19 MAN\_COLL 1.240615e-02  
## 20 RELJCT1 1.160944e-05  
## 21 RELJCT2 1.107517e-03  
## 22 TYP\_INT 3.259293e-04  
## 23 WRK\_ZONE 1.040211e-04  
## 24 REL\_ROAD 1.189199e-03  
## 25 LGT\_COND 4.553907e-04  
## 26 WEATHER 2.532999e-04  
## 27 SCH\_BUS 4.537089e-06  
## 28 NOT\_HOUR 5.877461e-04  
## 29 NOT\_MIN 1.102146e-03  
## 30 ARR\_HOUR 5.408250e-04  
## 31 ARR\_MIN 1.188102e-03  
## 32 HOSP\_HR 1.704245e-03  
## 33 HOSP\_MN 2.366024e-03  
## 34 CF1 6.464566e-04  
## 35 CF2 1.902533e-04  
## 36 CF3 1.369542e-04  
## 37 FATALS 2.569514e-01  
## 38 DRUNK\_DR 2.054720e-03

##### Using FSelector

All variables fcpp

#library(FSelector)  
weights\_fcpp <- information.gain(FATALS~., data = fcpp)  
row.names(weights\_fcpp)[order(weights\_fcpp, decreasing = TRUE)]

## [1] "ST\_CASE" "STATE" "TWAY\_ID" "LONGITUD" "TWAY\_ID2"   
## [6] "LATITUDE" "COUNTY" "CITY" "RD\_OWNER" "MILEPT"   
## [11] "NOT\_MIN" "ARR\_MIN" "ROUTE" "FUNC\_SYS" "ARR\_HOUR"   
## [16] "NOT\_HOUR" "HOSP\_MN" "RUR\_URB" "WEATHER1" "WEATHER"   
## [21] "HOSP\_HR" "PERSONS" "PERMVIT" "HARM\_EV" "TIMESTAMP"   
## [26] "CF1" "MONTH" "PERNOTMVIT" "PEDS" "RELJCT2"   
## [31] "RAIL" "MAN\_COLL" "CF2" "CF3" "NHS"   
## [36] "VE\_FORMS" "VE\_TOTAL" "LGT\_COND" "SP\_JUR" "TYP\_INT"   
## [41] "REL\_ROAD" "DRUNK\_DR" "WEATHER2" "MINUTE" "HOUR"   
## [46] "RELJCT1" "WRK\_ZONE" "PVH\_INVL" "DAY" "YEAR"   
## [51] "DAY\_WEEK" "SCH\_BUS"

Pre-processed variables fcpp16

#library(FSelector)  
weights\_fcpp16 <- information.gain(FATALS~., data = fcpp16)  
row.names(weights\_fcpp16)[order(weights\_fcpp16, decreasing = TRUE)]

## [1] "PERSONS" "PERMVIT" "PEDS" "HARM\_EV" "PERNOTMVIT"  
## [6] "MAN\_COLL" "VE\_FORMS" "VE\_TOTAL" "RUR\_URB" "DRUNK\_DR"   
## [11] "RD\_OWNER" "ROUTE" "MILEPT" "HOSP\_MN" "HOSP\_HR"   
## [16] "REL\_ROAD" "NHS" "FUNC\_SYS" "RELJCT2" "LONGITUD"   
## [21] "DAY\_WEEK" "SP\_JUR" "PVH\_INVL" "LATITUDE" "RELJCT1"   
## [26] "TYP\_INT" "WRK\_ZONE" "LGT\_COND" "WEATHER" "SCH\_BUS"   
## [31] "NOT\_HOUR" "NOT\_MIN" "ARR\_HOUR" "ARR\_MIN" "CF1"   
## [36] "CF2" "CF3"

#library(FSelector)  
weights\_fcpp16 <- information.gain(FATALS~., fcpp16)  
print(weights\_fcpp16)

## attr\_importance  
## VE\_TOTAL 0.0092735064  
## VE\_FORMS 0.0097152240  
## PVH\_INVL 0.0000000000  
## PEDS 0.0147743143  
## PERNOTMVIT 0.0137552504  
## PERMVIT 0.0414105155  
## PERSONS 0.0421227316  
## DAY\_WEEK 0.0003979821  
## NHS 0.0009320245  
## RUR\_URB 0.0020353414  
## FUNC\_SYS 0.0008940465  
## RD\_OWNER 0.0016879786  
## ROUTE 0.0016273312  
## MILEPT 0.0013753449  
## LATITUDE 0.0000000000  
## LONGITUD 0.0004274073  
## SP\_JUR 0.0003974751  
## HARM\_EV 0.0137801373  
## MAN\_COLL 0.0123075345  
## RELJCT1 0.0000000000  
## RELJCT2 0.0006782752  
## TYP\_INT 0.0000000000  
## WRK\_ZONE 0.0000000000  
## REL\_ROAD 0.0010863497  
## LGT\_COND 0.0000000000  
## WEATHER 0.0000000000  
## SCH\_BUS 0.0000000000  
## NOT\_HOUR 0.0000000000  
## NOT\_MIN 0.0000000000  
## ARR\_HOUR 0.0000000000  
## ARR\_MIN 0.0000000000  
## HOSP\_HR 0.0012659938  
## HOSP\_MN 0.0013439542  
## CF1 0.0000000000  
## CF2 0.0000000000  
## CF3 0.0000000000  
## DRUNK\_DR 0.0020195837

## Select top 15 variables  
subset\_fcpp16\_15 <- cutoff.k(weights\_fcpp16, 15)  
subset\_fcpp16\_15

## [1] "PERSONS" "PERMVIT" "PEDS" "HARM\_EV" "PERNOTMVIT"  
## [6] "MAN\_COLL" "VE\_FORMS" "VE\_TOTAL" "RUR\_URB" "DRUNK\_DR"   
## [11] "RD\_OWNER" "ROUTE" "MILEPT" "HOSP\_MN" "HOSP\_HR"

f15 <- as.simple.formula(subset\_fcpp16\_15, "FATALS")  
#row.names(f)[order(f, decreasing = TRUE)]  
print(f15)

## FATALS ~ PERSONS + PERMVIT + PEDS + HARM\_EV + PERNOTMVIT + MAN\_COLL +   
## VE\_FORMS + VE\_TOTAL + RUR\_URB + DRUNK\_DR + RD\_OWNER + ROUTE +   
## MILEPT + HOSP\_MN + HOSP\_HR  
## <environment: 0x0000000019471f68>

#library(FSelector)  
weights\_fcpp <- information.gain(FATALS~., fcpp)  
print(weights\_fcpp)

## attr\_importance  
## STATE 1.2630018636  
## ST\_CASE 1.3895666726  
## VE\_TOTAL 0.0038590897  
## VE\_FORMS 0.0039657822  
## PVH\_INVL 0.0000000000  
## PEDS 0.0067970047  
## PERNOTMVIT 0.0069187479  
## PERMVIT 0.0131702163  
## PERSONS 0.0134102077  
## COUNTY 0.2169627246  
## CITY 0.1885537893  
## DAY 0.0000000000  
## MONTH 0.0084342502  
## YEAR 0.0000000000  
## DAY\_WEEK 0.0000000000  
## HOUR 0.0011178225  
## MINUTE 0.0011343122  
## NHS 0.0042358030  
## RUR\_URB 0.0238234537  
## FUNC\_SYS 0.0374263698  
## RD\_OWNER 0.1194007392  
## ROUTE 0.0429665411  
## TWAY\_ID 1.0680525696  
## TWAY\_ID2 0.3945010648  
## MILEPT 0.0910941941  
## LATITUDE 0.3325902093  
## LONGITUD 0.5694119262  
## SP\_JUR 0.0031648160  
## HARM\_EV 0.0105534200  
## MAN\_COLL 0.0046241836  
## RELJCT1 0.0009658328  
## RELJCT2 0.0056126886  
## TYP\_INT 0.0028464428  
## WRK\_ZONE 0.0007217201  
## REL\_ROAD 0.0028366452  
## LGT\_COND 0.0037702708  
## WEATHER1 0.0236473302  
## WEATHER2 0.0015952917  
## WEATHER 0.0236473302  
## SCH\_BUS 0.0000000000  
## RAIL 0.0054467226  
## NOT\_HOUR 0.0368450705  
## NOT\_MIN 0.0686720109  
## ARR\_HOUR 0.0374137506  
## ARR\_MIN 0.0683217228  
## HOSP\_HR 0.0231986235  
## HOSP\_MN 0.0310036238  
## CF1 0.0095904082  
## CF2 0.0045479238  
## CF3 0.0045479238  
## DRUNK\_DR 0.0021235417  
## TIMESTAMP 0.0104822046

## Select top 10 variables  
subset\_fcpp\_10 <- cutoff.k(weights\_fcpp, 10)  
subset\_fcpp\_10

## [1] "ST\_CASE" "STATE" "TWAY\_ID" "LONGITUD" "TWAY\_ID2" "LATITUDE"  
## [7] "COUNTY" "CITY" "RD\_OWNER" "MILEPT"

f\_fcpp <- as.simple.formula(subset\_fcpp\_10, "FATALS")  
print(f\_fcpp)

## FATALS ~ ST\_CASE + STATE + TWAY\_ID + LONGITUD + TWAY\_ID2 + LATITUDE +   
## COUNTY + CITY + RD\_OWNER + MILEPT  
## <environment: 0x000000001980d298>

## Select top 15 variables  
subset\_fcpp\_15 <- cutoff.k(weights\_fcpp, 15)  
subset\_fcpp\_15

## [1] "ST\_CASE" "STATE" "TWAY\_ID" "LONGITUD" "TWAY\_ID2" "LATITUDE"  
## [7] "COUNTY" "CITY" "RD\_OWNER" "MILEPT" "NOT\_MIN" "ARR\_MIN"   
## [13] "ROUTE" "FUNC\_SYS" "ARR\_HOUR"

# The random.forest.importance function is used to rate the importance of each feature in the classification of the outcome, chd. The function returns a data frame containing the name of each attribute and the importance value based on the mean decrease in accuracy.

#library(ElemStatLearn)  
#library(FSelector)  
  
att.scores <- random.forest.importance(FATALS ~ ., fcpp16)  
att.scores

## attr\_importance  
## VE\_TOTAL 25.3389414  
## VE\_FORMS 24.9723888  
## PVH\_INVL 7.1366849  
## PEDS 18.3907909  
## PERNOTMVIT 22.9462374  
## PERMVIT 37.3902702  
## PERSONS 40.2136124  
## DAY\_WEEK 1.8221795  
## NHS 11.6359525  
## RUR\_URB 15.6436874  
## FUNC\_SYS 29.1389693  
## RD\_OWNER 20.4978635  
## ROUTE 25.9806050  
## MILEPT 29.1747615  
## LATITUDE 42.7421647  
## LONGITUD 35.8387633  
## SP\_JUR 8.2226761  
## HARM\_EV 24.4041784  
## MAN\_COLL 11.8910610  
## RELJCT1 6.0322241  
## RELJCT2 32.4847133  
## TYP\_INT 15.9720641  
## WRK\_ZONE -2.2087745  
## REL\_ROAD 34.4425547  
## LGT\_COND 16.3319194  
## WEATHER 6.3585851  
## SCH\_BUS 3.5264900  
## NOT\_HOUR 54.7633169  
## NOT\_MIN 52.1579297  
## ARR\_HOUR 48.7400061  
## ARR\_MIN 53.8744026  
## HOSP\_HR 63.1404348  
## HOSP\_MN 72.4433848  
## CF1 6.4917138  
## CF2 3.9695449  
## CF3 0.8421964  
## DRUNK\_DR 16.1087094

# The FSelector package offers several functions to choose the best features using the importance values returned by random.forest.importance.

# The cutoff.biggest.diff function automatically identifies the features which have a significantly higher importance value than other features.

# cutoff.k provides the k features with the highest importance values.

# Similarly, cutoff.k.percent returns k percent of the features with the highest importance values.

f1 <- cutoff.biggest.diff(att.scores)  
print(f1)

## [1] "HOSP\_MN"

f2 <- cutoff.k(att.scores, k = 15)  
f2

## [1] "HOSP\_MN" "HOSP\_HR" "NOT\_HOUR" "ARR\_MIN" "NOT\_MIN" "ARR\_HOUR"  
## [7] "LATITUDE" "PERSONS" "PERMVIT" "LONGITUD" "REL\_ROAD" "RELJCT2"   
## [13] "MILEPT" "FUNC\_SYS" "ROUTE"

f3 <- cutoff.k.percent(att.scores, 0.4)  
f3

## [1] "HOSP\_MN" "HOSP\_HR" "NOT\_HOUR" "ARR\_MIN" "NOT\_MIN" "ARR\_HOUR"  
## [7] "LATITUDE" "PERSONS" "PERMVIT" "LONGITUD" "REL\_ROAD" "RELJCT2"   
## [13] "MILEPT" "FUNC\_SYS" "ROUTE"

#library(FSelector)  
weights <- gain.ratio(FATALS ~., fcpp16)  
print(weights)

## attr\_importance  
## VE\_TOTAL 0.0135313356  
## VE\_FORMS 0.0111951275  
## PVH\_INVL 0.0000000000  
## PEDS 0.0269459617  
## PERNOTMVIT 0.0244516150  
## PERMVIT 0.0320713052  
## PERSONS 0.0325390244  
## DAY\_WEEK 0.0008518747  
## NHS 0.0013996792  
## RUR\_URB 0.0029380481  
## FUNC\_SYS 0.0018270147  
## RD\_OWNER 0.0014038546  
## ROUTE 0.0023935824  
## MILEPT 0.0020780312  
## LATITUDE 0.0000000000  
## LONGITUD 0.0006692534  
## SP\_JUR 0.0063706899  
## HARM\_EV 0.0108837566  
## MAN\_COLL 0.0118142779  
## RELJCT1 0.0000000000  
## RELJCT2 0.0015454434  
## TYP\_INT 0.0000000000  
## WRK\_ZONE 0.0000000000  
## REL\_ROAD 0.0016916107  
## LGT\_COND 0.0000000000  
## WEATHER 0.0000000000  
## SCH\_BUS 0.0000000000  
## NOT\_HOUR 0.0000000000  
## NOT\_MIN 0.0000000000  
## ARR\_HOUR 0.0000000000  
## ARR\_MIN 0.0000000000  
## HOSP\_HR 0.0011637213  
## HOSP\_MN 0.0012369101  
## CF1 0.0000000000  
## CF2 0.0000000000  
## CF3 0.0000000000  
## DRUNK\_DR 0.0033752379

weights = chi.squared(FATALS ~., fcpp16)  
print(weights)

## attr\_importance  
## VE\_TOTAL 0.13620022  
## VE\_FORMS 0.14197430  
## PVH\_INVL 0.00000000  
## PEDS 0.14462261  
## PERNOTMVIT 0.14082522  
## PERMVIT 0.28181519  
## PERSONS 0.28336912  
## DAY\_WEEK 0.02892377  
## NHS 0.04358537  
## RUR\_URB 0.06366587  
## FUNC\_SYS 0.04060545  
## RD\_OWNER 0.05562145  
## ROUTE 0.05625709  
## MILEPT 0.05150230  
## LATITUDE 0.00000000  
## LONGITUD 0.02887981  
## SP\_JUR 0.03185887  
## HARM\_EV 0.15402623  
## MAN\_COLL 0.17087192  
## RELJCT1 0.00000000  
## RELJCT2 0.03533233  
## TYP\_INT 0.00000000  
## WRK\_ZONE 0.00000000  
## REL\_ROAD 0.04569017  
## LGT\_COND 0.00000000  
## WEATHER 0.00000000  
## SCH\_BUS 0.00000000  
## NOT\_HOUR 0.00000000  
## NOT\_MIN 0.00000000  
## ARR\_HOUR 0.00000000  
## ARR\_MIN 0.00000000  
## HOSP\_HR 0.04925130  
## HOSP\_MN 0.05084214  
## CF1 0.00000000  
## CF2 0.00000000  
## CF3 0.00000000  
## DRUNK\_DR 0.07253120

result <- cfs(FATALS ~ ., fcpp16)  
f\_q <- as.simple.formula(result, "FATALS")  
print(f\_q)

## FATALS ~ PEDS + PERMVIT + PERSONS  
## <environment: 0x0000000062cc8200>

require(randomForest)

## Loading required package: randomForest

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

Y <- fcpp16$FATALS  
fit=randomForest(factor(Y)~., data=fcpp16)  
fit

##   
## Call:  
## randomForest(formula = factor(Y) ~ ., data = fcpp16)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 6  
##   
## OOB estimate of error rate: 0%  
## Confusion matrix:  
## 0 1 class.error  
## 0 31984 0 0  
## 1 0 2455 0