STAT149_project_extraData

Zecai Liang

4/24/2017

Assemble Data

```
##original data
data.train = read.csv('train.csv')
data.test = read.csv('test.csv')

##imputed data
train_X = read.csv("train_nafill_X.csv")
test_X = read.csv("test_nafill_X.csv")

## district data (colletec by John)
library(xlsx)
```

```
## Loading required package: rJava

## Loading required package: xlsxjars
```

```
hd.data = read.xlsx("149hddata.xlsx", sheetIndex = 1, header = TRUE)
# rename columns
colnames(hd.data) = c("hd", "hd.area", "hd.population", "hd.density", "hd.city")
# convert factor columns
hd.data$hd.city = factor(hd.data$hd.city)
```

```
head(hd.data)
```

```
##   hd hd.area hd.population hd.density hd.city
## 1  1   14.56         69128   4747.802       1
## 2  2    9.65         63544   6584.870       1
## 3  3   19.33         68147   3525.453       1
## 4  4    9.05         61240   6766.851       1
## 5  5   12.68         71316   5624.290       1
## 6  6   13.45         74553   5542.974       1
```

```
head(train_X)
```

```
##   gender cd hd age dbdistance vccdistance party  racename hsonly mrrg
## 1      M  7 31  36    1.97862     3.36181     U  Hispanic   25.4 63.4
## 2      F  6 38  55    3.21001     3.21633     U   Uncoded    7.9 97.8
## 3      F  2 53  24    1.93799     1.95190     U Caucasian   50.2  7.6
## 4      F  7 30  25    4.84987     4.76860     D Caucasian   38.0  8.5
## 5      M  5 19  22   10.25770    10.06500     R Caucasian   30.5 19.1
## 6      M  1  7  22    2.56360     2.65773     U Caucasian   32.0  7.5
##   chldprsnt cath evang nonchrst otherchrst days.since.reg
## 1      54.0 16.7  16.5     39.6       27.3            420
## 2      59.8 16.7  15.5     30.9       36.9            307
## 3      49.5 14.6  24.0     29.6       31.7            292
## 4      47.4 13.1  22.3     33.3       31.4            316
## 5      23.1 16.0  10.5     39.1       34.5            392
## 6      29.4 13.5  21.6     34.0       30.9            333
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

```
train_full_X = left_join(train_X, hd.data, by = c("hd"))
test_full_X = left_join(test_X, hd.data, by = c("hd"))
```

```
dim(train_full_X)
```

```
## [1] 118529     20
```

```
dim(test_full_X)
```

```
## [1] 39510    20
```

```
## convert "cd" and "hd" to factor
train_full_X$cd = factor(train_full_X$cd)
test_full_X$cd = factor(test_full_X$cd)

train_full_X$hd = factor(train_full_X$hd)
test_full_X$hd = factor(test_full_X$hd)
```

```
head(train_full_X)
```

```
##   gender cd hd age dbdistance vccdistance party  racename hsonly mrrg
## 1      M  7 31  36    1.97862     3.36181     U  Hispanic   25.4 63.4
## 2      F  6 38  55    3.21001     3.21633     U   Uncoded    7.9 97.8
## 3      F  2 53  24    1.93799     1.95190     U Caucasian   50.2  7.6
## 4      F  7 30  25    4.84987     4.76860     D Caucasian   38.0  8.5
## 5      M  5 19  22   10.25770    10.06500     R Caucasian   30.5 19.1
## 6      M  1  7  22    2.56360     2.65773     U Caucasian   32.0  7.5
##   chldprsnt cath evang nonchrst otherchrst days.since.reg hd.area
## 1      54.0 16.7  16.5     39.6       27.3            420   53.29
## 2      59.8 16.7  15.5     30.9       36.9            307   26.81
## 3      49.5 14.6  24.0     29.6       31.7            292   20.07
## 4      47.4 13.1  22.3     33.3       31.4            316  347.55
## 5      23.1 16.0  10.5     39.1       34.5            392 1327.13
## 6      29.4 13.5  21.6     34.0       30.9            333   68.27
##   hd.population hd.density hd.city
## 1        100635 1888.44061       1
## 2         70089 2614.28571       1
## 3         69496 3462.68062       1
## 4         82192  236.48971       0
## 5         81655   61.52751       0
## 6        101799 1491.12348       1
```

```
## add "voted" column to train data
train_full = cbind(voted = data.train$voted, train_full_X)
## add "id" column to test data
```

```
test_full = cbind(test_full_X, Id = data.test$Id)
```

```
head(train_full)
```

```
##   voted gender cd hd age dbdistance vccdistance party  racename hsonly
## 1     Y      M  7 31  36    1.97862     3.36181     U  Hispanic   25.4
## 2     Y      F  6 38  55    3.21001     3.21633     U   Uncoded    7.9
## 3     Y      F  2 53  24    1.93799     1.95190     U Caucasian   50.2
## 4     Y      F  7 30  25    4.84987     4.76860     D Caucasian   38.0
## 5     Y      M  5 19  22   10.25770    10.06500     R Caucasian   30.5
## 6     N      M  1  7  22    2.56360     2.65773     U Caucasian   32.0
##   mrrg chldprsnt cath evang nonchrst otherchrst days.since.reg hd.area
## 1 63.4      54.0 16.7  16.5     39.6       27.3            420   53.29
## 2 97.8      59.8 16.7  15.5     30.9       36.9            307   26.81
## 3  7.6      49.5 14.6  24.0     29.6       31.7            292   20.07
## 4  8.5      47.4 13.1  22.3     33.3       31.4            316  347.55
## 5 19.1      23.1 16.0  10.5     39.1       34.5            392 1327.13
## 6  7.5      29.4 13.5  21.6     34.0       30.9            333   68.27
##   hd.population hd.density hd.city
## 1        100635 1888.44061       1
## 2         70089 2614.28571       1
## 3         69496 3462.68062       1
## 4         82192  236.48971       0
## 5         81655   61.52751       0
## 6        101799 1491.12348       1
```

```
head(test_full)
```

```
##   gender cd hd age dbdistance vccdistance party  racename hsonly mrrg
## 1      M  2 52  30    2.21571     2.21750     L Caucasian   19.5 21.2
## 2      F  5 19  20    1.97090     1.78718     U Caucasian   39.7 20.2
## 3      M  4 44  56    2.13810     2.76109     R Caucasian   11.3 62.7
## 4      F  7 34  20    2.16572     2.92506     R Caucasian   32.8 11.6
## 5      F  6 41  26    4.81799     4.90072     D   Uncoded   10.2 14.7
## 6      F  2 11  45    2.07992     2.42841     D Caucasian   12.1 64.6
##   chldprsnt cath evang nonchrst otherchrst days.since.reg hd.area
## 1      25.3  9.8  16.6     45.2       28.4            393   42.07
## 2      29.1 12.0  14.4     41.4       32.2            668 1327.13
## 3      41.3 14.8  14.7     36.0       34.6            606  205.14
## 4      33.1 14.5  10.3     44.6       30.6            565   14.51
## 5      22.4  8.2  18.4     43.5       29.9            336   12.30
## 6      64.7 12.6  11.8     41.2       34.5            395   45.20
##   hd.population hd.density hd.city Id
## 1         80636 1916.71024       1  1
## 2         81655   61.52751       0  2
## 3        111170  541.92259       0  3
## 4         72738 5012.95658       1  4
## 5         74088 6023.41463       1  5
## 6         80189 1774.09292       1  6
```

```
## save to local files
write.csv(train_full, "train_full.csv", row.names = FALSE)
write.csv(test_full, "test_full.csv", row.names = FALSE)
```