Prediction Models

Jingwen Yin jy2786

April 24, 2017

GPA

```
source("../lib/modelFunc.R")
data.filtered <- read.csv('../data/NAreplaced.csv') #4242 1388
select <- read.csv('../data/Updated_Features/gpa_features.csv')
select.idx<-colnames(data.filtered) %in% as.character(select$Codes)
data.filtered <- data.filtered[,select.idx]

label <- read.csv('../data/train.csv')
label<-label[!is.na(label$gpa),]
Index<-as.numeric(rownames(data.filtered))%in% label$challengeID

data.train<-data.filtered[Index,]
data.train<-as.data.frame(data.train)
data.train<-cbind(label$gpa, data.train)
colnames(data.train)[1]<-"gpa"

# create training and test data set
set.seed(123)
train.index <- sample(1:nrow(data.train),800,replace = F)
train <- data.train[train.index,] #800*64
test <- data.train[-train.index,] #214*64

y<-train[,1]
model_selection_con(train[,-1], test, y)
```

```
##                          Method Test.Error
## 1             Linear Regression     0.4082
## 2                     Full tree     0.4303
## 3                   Pruned tree     0.4617
## 4                 Random Forest     0.3886
## 5  Conditional inference trees     0.4476
## 6                    gamboostLSS     0.3991
## 7             Gradient Boosting     0.3858
## 8        Support Vector Machine     0.4047
## 9                          LM+RF     0.3915
## 10                        SVM+RF     0.3917
```

Grit

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/grit_features.csv')
select.idx<-colnames(data.filtered) %in% as.character(select$Codes)
data.filtered <- data.filtered[,select.idx]
```

```
label <- read.csv('../data/train.csv')
label<-label[!is.na(label$grit),]
Index<-as.numeric(rownames(data.filtered))%in% label$challengeID

data.train<-data.filtered[Index,]
data.train<-as.data.frame(data.train)
data.train<-cbind(label$grit, data.train)
colnames(data.train)[1]<-"grit"

# create training and test data set
train.index <- sample(1:nrow(data.train),800,replace = F)
train <- data.train[train.index,] #800*64
test <- data.train[-train.index,] #214*64

y<-train[,1]
model_selection_con(train[,-1], test, y)
```

```
##                         Method Test.Error
## 1            Linear Regression     0.2213
## 2                    Full tree     0.2325
## 3                  Pruned tree     0.2259
## 4                Random Forest     0.2246
## 5  Conditional inference trees     0.2252
## 6                   gamboostLSS     0.2209
## 7            Gradient Boosting     0.2210
## 8       Support Vector Machine     0.2211
## 9                         LM+RF     0.2184
## 10                       SVM+RF     0.2176
```

materialHardship

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/materialHardship_features.csv')
select.idx<-colnames(data.filtered) %in% as.character(select$Codes)
data.filtered <- data.filtered[,select.idx]

label <- read.csv('../data/train.csv')
label<-label[!is.na(label$materialHardship),]
Index<-as.numeric(rownames(data.filtered))%in% label$challengeID

data.train<-data.filtered[Index,]
data.train<-as.data.frame(data.train)
data.train<-cbind(label$materialHardship, data.train)
colnames(data.train)[1]<-"materialHardship"

# create training and test data set
train.index <- sample(1:nrow(data.train),800,replace = F)
train <- data.train[train.index,] #800*64
test <- data.train[-train.index,] #214*64

y<-train[,1]
```

```
model_selection_con(train[,-1], test, y)
```

```
##                          Method Test.Error
## 1             Linear Regression     0.0194
## 2                     Full tree     0.0233
## 3                   Pruned tree     0.0214
## 4                 Random Forest     0.0194
## 5  Conditional inference trees     0.0222
## 6                   gamboostLSS     0.0441
## 7             Gradient Boosting     0.0197
## 8       Support Vector Machine     0.0212
## 9                         LM+RF     0.0256
## 10                       SVM+RF     0.0191
```

eviction

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/eviction_features.csv')
select.idx<-colnames(data.filtered) %in% as.character(select$Codes)
data.filtered <- data.filtered[,select.idx]

label <- read.csv('../data/train.csv')
label<-label[!is.na(label$eviction),]
Index<-as.numeric(rownames(data.filtered))%in% label$challengeID

data.train<-data.filtered[Index,]
data.train<-as.data.frame(data.train)
data.train<-cbind(label$eviction, data.train)
colnames(data.train)[1]<-"eviction"

# create training and test data set
train.index <- sample(1:nrow(data.train),800,replace = F)
train <- data.train[train.index,] #800*64
test <- data.train[-train.index,] #214*64

y<-factor(train[,1])
model_selection_cat(train[,-1], test, y)
```

```
##                          Method Test.Error
## 1                           glm     0.0759
## 2                     Full tree     0.0698
## 3                   Pruned tree     0.0577
## 4                 Random Forest     0.0577
## 5  Conditional inference trees     0.0577
## 6             Gradient Boosting     0.0637
## 7       Support Vector Machine     0.0577
## 8                         C5.0     0.0577
## 9                          LDA     0.0789
## 10                         KNN     0.0577
```

layoff

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/layoff_features.csv')
select.idx<-colnames(data.filtered) %in% as.character(select$Codes)
data.filtered <- data.filtered[,select.idx]

label <- read.csv('../data/train.csv')
label<-label[!is.na(label$layoff),]
Index<-as.numeric(rownames(data.filtered))%in% label$challengeID

data.train<-data.filtered[Index,]
data.train<-as.data.frame(data.train)
data.train<-cbind(label$layoff, data.train)
colnames(data.train)[1]<-"layoff"

# create training and test data set
train.index <- sample(1:nrow(data.train),800,replace = F)
train <- data.train[train.index,] #800*64
test <- data.train[-train.index,] #214*64

y<-factor(train[,1])
model_selection_cat(train[,-1], test, y)
```

```
##                          Method Test.Error
## 1                           glm     0.2138
## 2                     Full tree     0.2683
## 3                   Pruned tree     0.2138
## 4                 Random Forest     0.2306
## 5  Conditional inference trees     0.2138
## 6             Gradient Boosting     0.2222
## 7       Support Vector Machine     0.2117
## 8                           C5.0     0.2138
## 9                           LDA     0.2138
## 10                          KNN     0.2243
```

jobTraining

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/jobTraining_features.csv')
select.idx<-colnames(data.filtered) %in% as.character(select$Codes)
data.filtered <- data.filtered[,select.idx]

label <- read.csv('../data/train.csv')
label<-label[!is.na(label$jobTraining),]
Index<-as.numeric(rownames(data.filtered))%in% label$challengeID

data.train<-data.filtered[Index,]
data.train<-as.data.frame(data.train)
data.train<-cbind(label$jobTraining, data.train)
colnames(data.train)[1]<-"jobTraining"
```

```r
# create training and test data set
train.index <- sample(1:nrow(data.train),800,replace = F)
train <- data.train[train.index,] #800*64
test <- data.train[-train.index,] #214*64

y<-factor(train[,1])
model_selection_cat(train[,-1], test, y)
```

```
##                          Method Test.Error
## 1                           glm     0.2163
## 2                     Full tree     0.2632
## 3                   Pruned tree     0.2224
## 4                 Random Forest     0.2315
## 5  Conditional inference trees     0.2224
## 6             Gradient Boosting     0.2300
## 7       Support Vector Machine     0.2239
## 8                           C5.0     0.2224
## 9                           LDA     0.2194
## 10                          KNN     0.2572
```