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<-na.omit(label)
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     0.3498
## 2                    Full tree     0.3806
## 3                  Pruned tree     0.4065
## 4                Random Forest     0.3313
## 5  Conditional inference trees     0.3814
## 6                  gamboostLSS     0.3324
## 7            Gradient Boosting     0.3216
## 8       Support Vector Machine     0.3436
## 9                        LM+RF     0.3327
## 10                      SVM+RF     0.3298
```

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$Codes)
data.filtered <- data.filtered[,select.idx]
data.train<-data.filtered[Index,]
data.train<-as.data.frame(data.train)
data.train<-cbind(label$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     0.2077
## 2                    Full tree     0.2347
## 3                  Pruned tree     0.2212
## 4                Random Forest     0.2020
## 5  Conditional inference trees     0.2200
## 6                  gamboostLSS     0.2042
## 7            Gradient Boosting     0.2096
## 8       Support Vector Machine     0.2064
## 9                        LM+RF     0.2023
## 10                      SVM+RF     0.2005
```

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]

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     0.0252
## 2                    Full tree     0.0238
## 3                  Pruned tree     0.0257
## 4                Random Forest     0.0232
## 5  Conditional inference trees     0.0276
## 6                 gamboostLSS     0.0573
## 7           Gradient Boosting     0.0239
## 8      Support Vector Machine     0.0290
## 9                       LM+RF     0.0259
## 10                     SVM+RF     0.0251
```

eviction

```r
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]

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     0.0748
## 2                     Full tree     0.0654
## 3                   Pruned tree     0.0654
## 4                 Random Forest     0.0654
## 5   Conditional inference trees     0.0654
## 6             Gradient Boosting     0.0701
## 7        Support Vector Machine     0.0654
## 8                          C5.0     0.0794
## 9                           LDA     0.0794
## 10                          KNN     0.0654
```

layoff

```r
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]

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

```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     0.2009
## 2                    Full tree     0.2150
## 3                  Pruned tree     0.1963
## 4                Random Forest     0.2009
## 5  Conditional inference trees     0.1963
## 6            Gradient Boosting     0.1963
## 7       Support Vector Machine     0.1963
## 8                         C5.0     0.1963
## 9                          LDA     0.2009
## 10                         KNN     0.2196
```

jobTraining

```r
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]

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

# 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     0.2243
## 2                    Full tree     0.2477
## 3                  Pruned tree     0.2290
## 4                Random Forest     0.2664
## 5  Conditional inference trees     0.2290
## 6            Gradient Boosting     0.2383
## 7       Support Vector Machine     0.2243
## 8                         C5.0     0.2944
## 9                          LDA     0.2243
## 10                         KNN     0.2664
```