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')
data.filtered <- data.filtered[,select$Codes] # 4242*64

label <- read.csv('../data/train.csv')
label<-na.omit(label)
Index<-data.filtered$challengeID %in% label$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     0.3525
## 2                     Full tree     0.3871
## 3                   Pruned tree     0.3645
## 4                 Random Forest     0.3386
## 5  Conditional inference trees     0.3871
## 6                   gamboostLSS     0.3360
## 7             Gradient Boosting     0.3357
## 8       Support Vector Machine     0.3408
## 9                         LM+RF     0.3396
## 10                       SVM+RF     0.3340
```

Grit

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/grit_features.csv')
data.filtered <- data.filtered[,select$Codes]

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     0.2227
## 2                     Full tree     0.2238
## 3                   Pruned tree     0.2259
## 4                 Random Forest     0.2322
## 5  Conditional inference trees     0.2259
## 6                   gamboostLSS     0.2202
## 7             Gradient Boosting     0.2224
## 8        Support Vector Machine     0.2289
## 9                         LM+RF     0.2247
## 10                       SVM+RF     0.2246
```

materialHardship

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/materialHardship_features.csv')
data.filtered <- data.filtered[,select$Codes]

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     0.0216
## 2                     Full tree     0.0217
## 3                   Pruned tree     0.0209
## 4                 Random Forest     0.0195
## 5  Conditional inference trees     0.0198
## 6                   gamboostLSS     0.0705
## 7             Gradient Boosting     0.0192
## 8        Support Vector Machine     0.0223
```

```
## 9                        LM+RF     0.0198
## 10                      SVM+RF     0.0199
```

eviction

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/eviction_features.csv')
data.filtered <- data.filtered[,select$Codes]

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     0.0701
## 2                   Full tree     0.0607
## 3                  Pruned tree     0.0654
## 4                Random Forest     0.0654
## 5  Conditional inference trees     0.0654
## 6            Gradient Boosting     0.0654
## 7       Support Vector Machine     0.0654
## 8                         C5.0     0.0654
## 9                          LDA     0.0748
## 10                         KNN     0.0654
```

layoff

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/layoff_features.csv')
data.filtered <- data.filtered[,select$Codes]

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     0.2196
## 2                     Full tree     0.2243
## 3                   Pruned tree     0.2243
## 4                 Random Forest     0.2243
## 5  Conditional inference trees     0.2243
## 6             Gradient Boosting     0.2243
## 7        Support Vector Machine     0.2243
## 8                            C5.0     0.2243
## 9                            LDA     0.2196
## 10                           KNN     0.2383
```

jobTraining

```
data.filtered <- read.csv('../data/NAreplaced.csv')
select <- read.csv('../data/Updated_Features/jobTraining_features.csv')
data.filtered <- data.filtered[,select$Codes]

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     0.2617
## 2                     Full tree     0.2430
## 3                   Pruned tree     0.2383
## 4                 Random Forest     0.2290
## 5  Conditional inference trees     0.2383
## 6             Gradient Boosting     0.2336
## 7        Support Vector Machine     0.2383
## 8                            C5.0     0.2944
## 9                            LDA     0.2570
## 10                           KNN     0.2477
```