Week1

Jocelyn Jin

November 9, 2017

prepare data

```
tests<-read.csv("T_FRUDSAB.csv")
test<-tests[,c("patdeid","VISIT", "UDS011")]
reg.data<-data.frame(unique(test$patdeid))
reg.data$week1<-rep(NA, nrow(reg.data))
reg.data$week2<-rep(NA, nrow(reg.data))
reg.data$week3<-rep(NA, nrow(reg.data))
reg.data$week4<-rep(NA, nrow(reg.data))
reg.data$week21<-rep(NA, nrow(reg.data))
reg.data$week22<-rep(NA, nrow(reg.data))
reg.data$week23<-rep(NA, nrow(reg.data))
reg.data$week24<-rep(NA, nrow(reg.data))
i<-1
for(i in 1:nrow(reg.data)){
  reg.data$week1[i]<-ifelse(length(test[test$patdeid==i&test$VISIT=="WK1","UDS011"])!=0,
                            test[test$patdeid==i&test$VISIT=="WK1", "UDS011"], NA)
  reg.data$week2[i]<-ifelse(length(test[test$patdeid==i&test$VISIT=="WK2","UDS011"])!=0,
                            test[test$patdeid==i&test$VISIT=="WK2", "UDS011"], NA)
  reg.data$week3[i]<-ifelse(length(test[test$patdeid==i&test$VISIT=="WK3","UDS011"])!=0,
                            test[test$patdeid==i&test$VISIT=="WK3", "UDS011"], NA)
  reg.data$week4[i]<-ifelse(length(test[test$patdeid==i&test$VISIT=="WK4","UDS011"])!=0,
                            test[test$patdeid==i&test$VISIT=="WK4", "UDS011"], NA)
  reg.data$week21[i]<-ifelse(length(test[test$patdeid==i&test$VISIT=="WK21","UDS011"])!=0,
                            test[test$patdeid==i&test$VISIT=="WK21", "UDS011"], NA)
  reg.data$week22[i]<-ifelse(length(test[test$patdeid==i&test$VISIT=="WK22","UDS011"])!=0,
                            test[test$patdeid==i&test$VISIT=="WK22", "UDS011"], NA)
  reg.data$week23[i]<-ifelse(length(test[test$patdeid==i&test$VISIT=="WK23","UDS011"])!=0,
                            test[test$patdeid==i&test$VISIT=="WK23", "UDS011"], NA)
  reg.data$week24[i]<-ifelse(length(test[test$patdeid==i&test$VISIT=="WK24","UDS011"])!=0,
                            test[test$patdeid==i&test$VISIT=="WK24", "UDS011"], NA)
}
```

lable the sample

```
reg.data$lab<-rep(NA, nrow(reg.data))
i<-2
for(i in 1:nrow(reg.data)){
  reg.data$lab[i]<-ifelse(sum(is.na(reg.data[i,6:9])), 1,
          ifelse(sum(reg.data[i,6:9])!=0,1,0))
  }
table(reg.data$lab)
```

```
##
```

```
##    0    1
##  294 1623
```

EDA

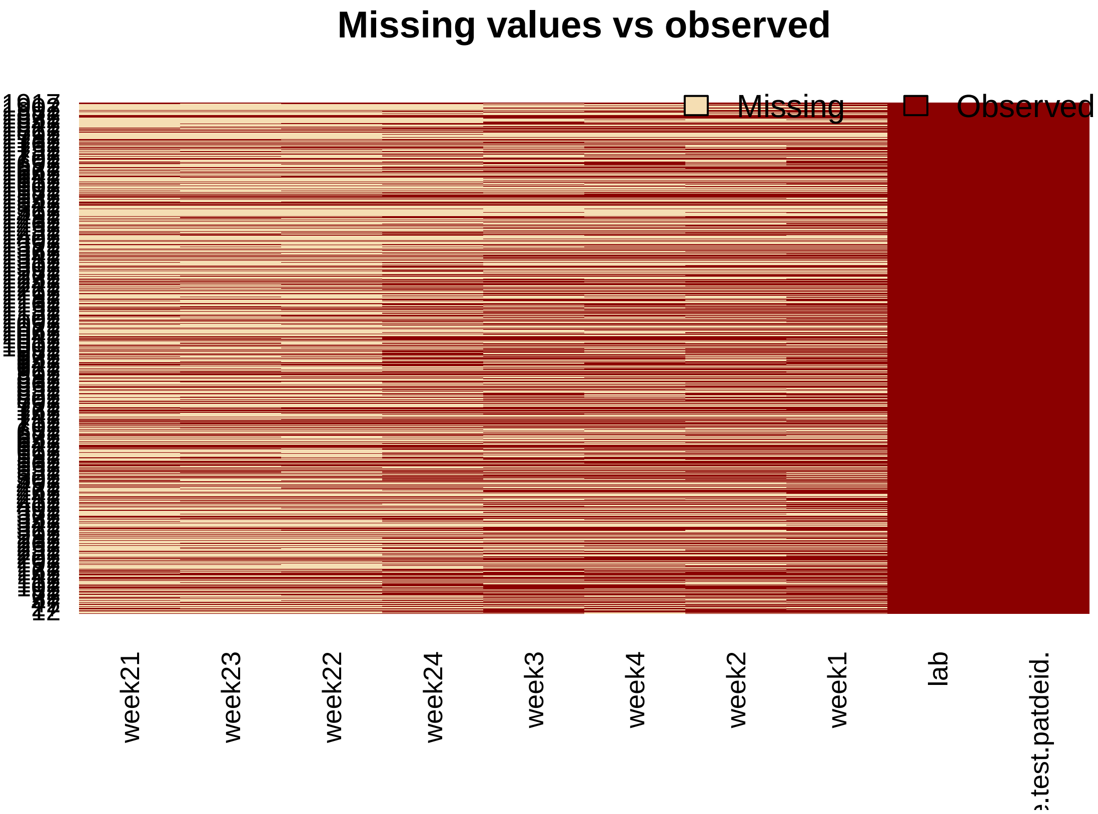
```
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 3.4.2
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.4, built: 2015-12-05)
## ## Copyright (C) 2005-2017 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
missmap(reg.data, main = "Missing values vs observed")
```



##logistic regression

```
sapply(reg.data, function(x) sum(is.na(x)))
```

```
## unique.test.patdeid.                week1                week2
##                    0                  853                  903
##                week3                week4               week21
##                  973                  952                 1268
##               week22               week23               week24
##                 1254                 1267                 1058
##                  lab
##                    0
```

```
data <- reg.data[-which(apply(reg.data[,2:5],1,function(x)all(is.na(x)))),]
#data partition
data$group<-sample(c(1,1,1,2), size=nrow(data), replace = TRUE)
train <- data[data$group==1,]
test <- data[data$group==2,]
#model fitting
model <- glm(lab~week1+week2+week3+week4,family=binomial(link='logit'),data=train)
summary(model)
```

```
##
## Call:
## glm(formula = lab ~ week1 + week2 + week3 + week4, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7867  -1.3107   0.6731   0.9039   3.1451
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3079     0.1120   2.749  0.00598 **
## week1         0.4611     0.1577   2.923  0.00347 **
## week2         0.2141     0.1315   1.628  0.10351
## week3         0.0103     0.1236   0.083  0.93363
## week4         0.3762     0.1580   2.382  0.01723 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 780.71  on 613  degrees of freedom
## Residual deviance: 747.42  on 609  degrees of freedom
##   (237 observations deleted due to missingness)
## AIC: 757.42
##
## Number of Fisher Scoring iterations: 4
```

```
predict <- predict(model, newdata=train, type = 'response')
```

ROC Curve

```
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.4.2
```

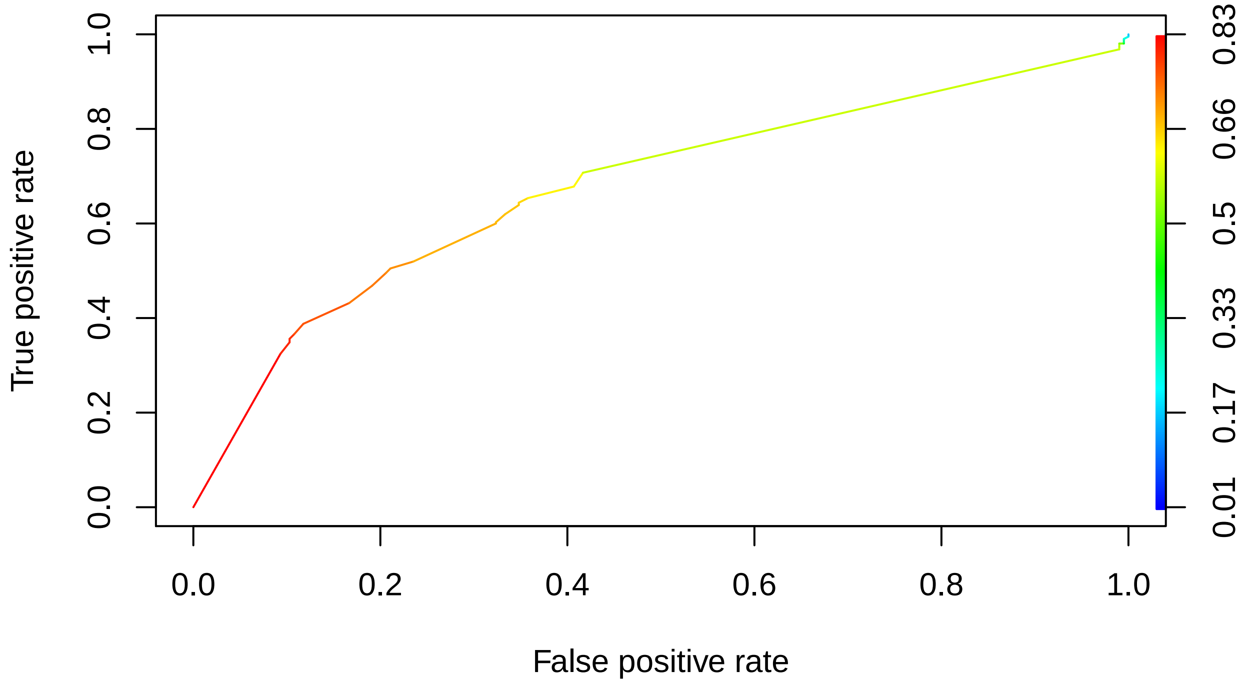
```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.4.2
```

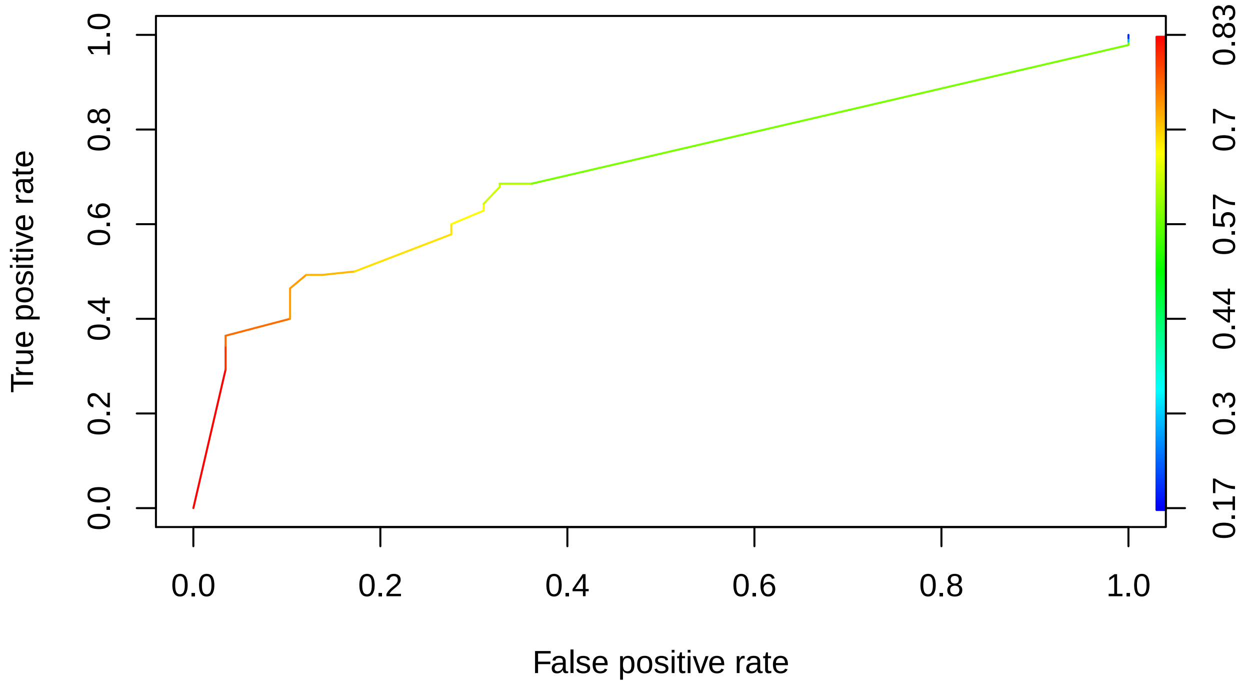
```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
#train data
ROCRpred <- prediction(predict, train$lab)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
```



```
#test data
tpredict<-predict(model, newdata=test, type = 'response')
tROCRpred <- prediction(tpredict, test$lab)
tROCRperf <- performance(tROCRpred, 'tpr','fpr')
plot(tROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
```



k fold cross validation

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.2

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
#don't remove missing values
data$week1[is.na(data$week1)] <- mean(data$week1,na.rm=T)
data$week2[is.na(data$week2)] <- mean(data$week2,na.rm=T)
data$week3[is.na(data$week3)] <- mean(data$week3,na.rm=T)
data$week4[is.na(data$week4)] <- mean(data$week4,na.rm=T)
sapply(data, function(x) sum(is.na(x)))
```

```
## unique.test.patdeid.                week1                week2
##                    0                    0                    0
##                week3                week4               week21
##                    0                    0                  478
##               week22               week23               week24
##                  464                  477                  295
##                  lab                group
##                     0                    0
```

```
ctrl <- trainControl(method = "cv", number = 10, savePredictions = T)
data$lab<-as.factor(data$lab)
glm_fit <- train(lab~week1+week2+week3+week4,
                    data = data,
                    method = "glm",
                    family=binomial(link='logit'),
                    trControl = ctrl)
glm_fit
```

```
## Generalized Linear Model
##
## 1127 samples
##    4 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1014, 1015, 1013, 1014, 1015, 1015, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7151775  -0.04218621
```

```
glm_fit$finalModel
```

```
##
## Call:  NULL
##
## Coefficients:
## (Intercept)        week1        week2        week3        week4
##    0.831484     0.180365     0.004552     0.142134     0.321666
##
## Degrees of Freedom: 1126 Total (i.e. Null);  1122 Residual
## Null Deviance:       1294
## Residual Deviance: 1268  AIC: 1278
```

```
head(glm_fit$pred)
```

```
##   pred obs rowIndex parameter Resample
## 1    1   1        6      none   Fold01
## 2    1   0       15      none   Fold01
## 3    1   0       20      none   Fold01
## 4    1   1       21      none   Fold01
## 5    1   1       23      none   Fold01
## 6    1   1       34      none   Fold01
```