Untitled

Sergio Grueso

11/10/2019

# Index:

# - Intro

# - Method:

# 1) Data analysis and visualization

# 2) Moddeling Aproach

# - Results: summary of all the accuracies

# - Conclusions:

# 

# Download al load the libraries and data of the longitudinal and cross sectional studies (cross sectional study will only be used as a test set). All can be found in <https://github.com/sgrueso/MRI.git>.

if(!require(tidyverse)) install.packages("tidyverse")

if(!require(caret)) install.packages("caret")

if(!require(dslabs)) install.packages("dslabs")

if(!require(dbplyr)) install.packages("dbplyr")

if(!require(ggplot2)) install.packages("ggplot2")  
if(!require(Rborist)) install.packages("Rborist")

if(!require(randomForest)) install.packages("randomForest")

set.seed(1, sample.kind = "Rounding")

longitudinal <- read.csv("../MRI/oasis\_longitudinal.csv", header = TRUE, stringsAsFactors = FALSE)

# Summary of the longitudinal dataset

summary(longitudinal)

## Subject.ID MRI.ID Group Visit   
## Length:373 Length:373 Length:373 Min. :1.000   
## Class :character Class :character Class :character 1st Qu.:1.000   
## Mode :character Mode :character Mode :character Median :2.000   
## Mean :1.882   
## 3rd Qu.:2.000   
## Max. :5.000   
##   
## MR.Delay M.F Hand Age   
## Min. : 0.0 Length:373 Length:373 Min. :60.00   
## 1st Qu.: 0.0 Class :character Class :character 1st Qu.:71.00   
## Median : 552.0 Mode :character Mode :character Median :77.00   
## Mean : 595.1 Mean :77.01   
## 3rd Qu.: 873.0 3rd Qu.:82.00   
## Max. :2639.0 Max. :98.00   
##   
## EDUC SES MMSE CDR   
## Min. : 6.0 Min. :1.00 Min. : 4.00 Min. :0.0000   
## 1st Qu.:12.0 1st Qu.:2.00 1st Qu.:27.00 1st Qu.:0.0000   
## Median :15.0 Median :2.00 Median :29.00 Median :0.0000   
## Mean :14.6 Mean :2.46 Mean :27.34 Mean :0.2909   
## 3rd Qu.:16.0 3rd Qu.:3.00 3rd Qu.:30.00 3rd Qu.:0.5000   
## Max. :23.0 Max. :5.00 Max. :30.00 Max. :2.0000   
## NA's :19 NA's :2   
## eTIV nWBV ASF   
## Min. :1106 Min. :0.6440 Min. :0.876   
## 1st Qu.:1357 1st Qu.:0.7000 1st Qu.:1.099   
## Median :1470 Median :0.7290 Median :1.194   
## Mean :1488 Mean :0.7296 Mean :1.195   
## 3rd Qu.:1597 3rd Qu.:0.7560 3rd Qu.:1.293   
## Max. :2004 Max. :0.8370 Max. :1.587   
##

## 

## Data analysis:

# Hand variable is not relevant because it has the same values for all subjects. Hand is droped.

longitudinal$Hand <- NULL

# M.F variable name is changed to “Sex” to be more descriptive

names(longitudinal)[names(longitudinal) == "M.F"] <- "Sex"

# Quick easy search for null values

sum(is.na(longitudinal))

## [1] 21

# 19 observations have one missing value from SES variable and 2 observations have missing values from SES and MMSE column. Subjects with 2 missing values are dropped and any missing value is replaced by the median of that variable. Median is used instead of mean because SES is a variable with integer values from 1 to 5 that indicates socioeconomical status and we do not want decimals values.

ind <- which(is.na(longitudinal$MMSE))  
longitudinal <- longitudinal[-ind, ]  
  
ind <- which(is.na(longitudinal$SES))  
longitudinal$SES[ind] <- median(longitudinal$SES[-ind])

# Group variable is the variable we will want to predict so demented is changed to 1 and nondemented to 0. This is in order to facilitate predictions, if the algorithm predicts anything greater than 0.5 (>0.5) the prediction will be “1” meaning that pacient has or is in risk of Alzheimer.

longitudinal$Group <- ifelse(longitudinal$Group=="Demented", 1, 0)  
longitudinal$Group <- as.factor(longitudinal$Group)

# correlations between variables

longitudinal %>% select(MR.Delay, Age, EDUC, MMSE, CDR, eTIV, nWBV, ASF) %>% cor()

## MR.Delay Age EDUC MMSE CDR  
## MR.Delay 1.00000000 0.20550115 0.05354613 0.06584393 -0.06729902  
## Age 0.20550115 1.00000000 -0.02859911 0.05561218 -0.02514668  
## EDUC 0.05354613 -0.02859911 1.00000000 0.19488436 -0.14561447  
## MMSE 0.06584393 0.05561218 0.19488436 1.00000000 -0.68651948  
## CDR -0.06729902 -0.02514668 -0.14561447 -0.68651948 1.00000000  
## eTIV 0.12451322 0.04143405 0.25088721 -0.03208381 0.04245037  
## nWBV -0.10582013 -0.51825368 -0.01118340 0.34191241 -0.35047275  
## ASF -0.12997991 -0.03395663 -0.23448371 0.04005216 -0.05376694  
## eTIV nWBV ASF  
## MR.Delay 0.12451322 -0.1058201 -0.12997991  
## Age 0.04143405 -0.5182537 -0.03395663  
## EDUC 0.25088721 -0.0111834 -0.23448371  
## MMSE -0.03208381 0.3419124 0.04005216  
## CDR 0.04245037 -0.3504728 -0.05376694  
## eTIV 1.00000000 -0.2099807 -0.98912256  
## nWBV -0.20998073 1.0000000 0.21391424  
## ASF -0.98912256 0.2139142 1.00000000

# ASF is the computed scaling factor that transforms native-space brain and skull to the atlas target (i.e., the determinant of the transform matrix) and eTIV (estimated total intracranial volume in cm3) is the automated estimate of total intracranial volume in native space derived from the ASF (Buckner et al. 2004). That is why both variables are so highly correlated.

# Other significanly correlated (c < 0.20) variables are:

# - Age and MR.Delay

# - Age and nWBV

# - Education with eTIV and ASF

# - nWBV with eTIV and ASF

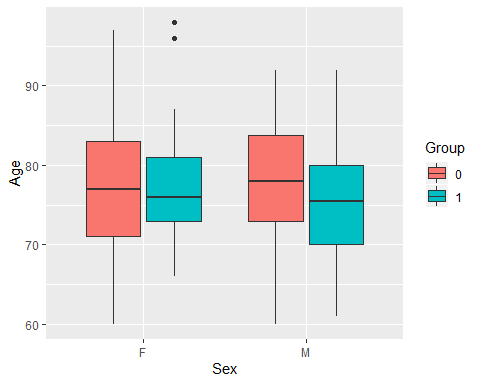
# - CDR and MMSE

## 

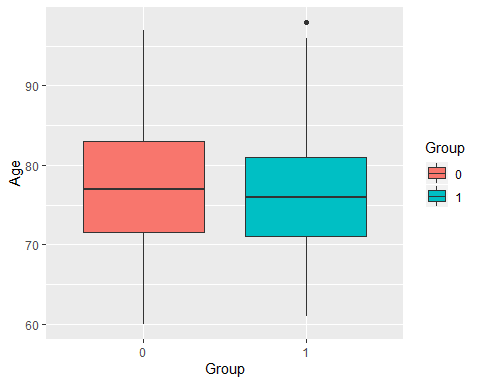
## Data visualization

# Data visualization analysis can be started by observing the Group, Age and Sex distribution

longitudinal %>% ggplot(aes(Sex, Age, fill= Group)) +  
 geom\_boxplot()

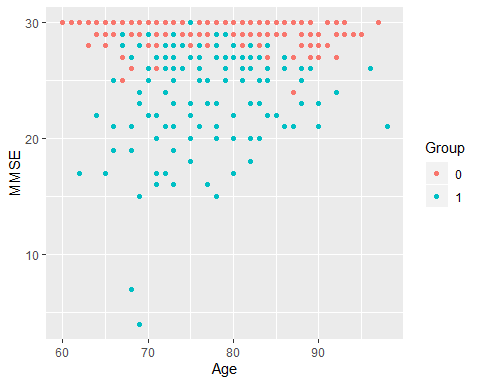


longitudinal %>% ggplot(aes(Group, Age, fill = Group)) +  
 geom\_boxplot()



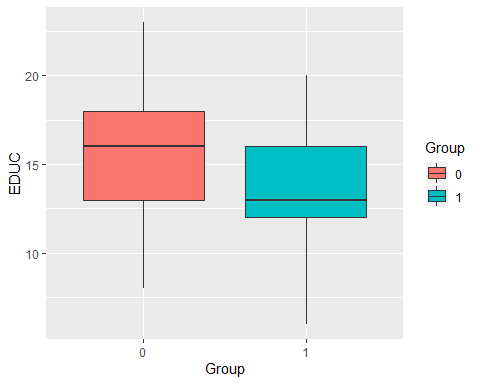
# Results on the test MMSE plotted against Age and colored depending on the Group.

longitudinal %>% ggplot(aes(Age, MMSE, col = Group)) +  
 geom\_point()



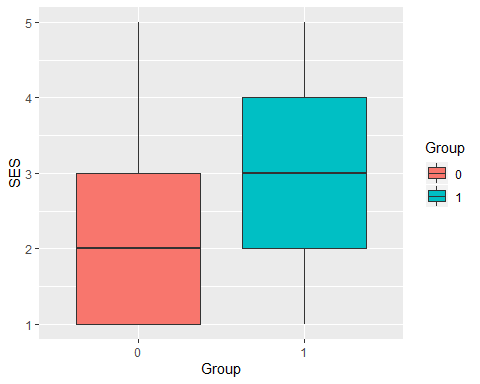
# Comparison between Education and Group

longitudinal %>% ggplot(aes(Group, EDUC, fill = Group)) + geom\_boxplot()



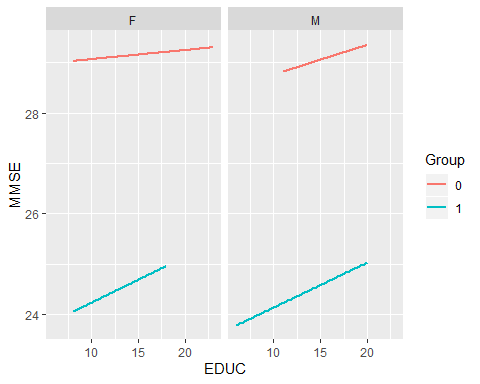
# SES vs Group

longitudinal %>% ggplot(aes(Group, SES, fill = Group)) + geom\_boxplot()



# As previous plots have shown, Education and MMSE have a negative correlation with Dementia. Therefore a final plot of MMSE against education is suggested with the expectation to find a positive correlation bewtween these 2 variables, and also, independently of the value in the variables Sex and Group.

longitudinal %>% ggplot(aes(EDUC, MMSE, col = Group)) + geom\_smooth(method = "lm", se = FALSE) + facet\_wrap(~Sex)



In summary, patients with Alzheimer (dementia) compared with non-demented subjects; tend to be less educated, in a lower social economic status and with lower punctuations in MMSE.

Age is normally the principal risk factor but in this study the mean age of the group with dementia is actually lower. That is probably due to the fact that the expectancy of life in people with dementia is much lower. Therefore we do not see a lot of people arrive to mid to late 90s with Alzheimer’s disease.

# BUILDING THE MODEL

# Train and test set partitions are made:

set.seed(1, sample.kind="Rounding")

## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding'  
## sampler used

test\_index <- createDataPartition(y = longitudinal$Group, times = 1, p = 0.2, list = FALSE)  
train\_set <- longitudinal[-test\_index,]  
test\_set <- longitudinal[test\_index,]

# Naive Baseline Model:

# The obvious way of predicting dementia seems to be MMSE. A value lower or equal than 27 predicts dementia.

y\_hat <- ifelse(test\_set$MMSE <= 27, 1, 0)  
  
mean(y\_hat == test\_set$Group)

## [1] 0.84

# The model has a 0.84 accuracy with only one parameter, that is unsual and very high accuracy. But can I improve the accuracy if the model includes more parameters like Age, Education or Social Economic Status?.

# MMSE is a test and includes scales that have into account Age and level of education of the subject. That is represented in our study in the variables Age and EDUC so the next model must include both variables.

set.seed(1, sample.kind="Rounding")

## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding'  
## sampler used

model1 <- train(Group ~ Age + EDUC + MMSE,  
 method = "glm",  
 data = train\_set)

confusionMatrix(predict(model1, test\_set), test\_set$Group)$overall["Accuracy"]

## Accuracy   
## 0.8666667

# The model is improved by adding both variables. Can more variables improve more the accurac?

set.seed(1, sample.kind="Rounding")

model2 <- train(Group ~ Age + EDUC + MMSE + SES + eTIV + nWBV,  
 method = "glm",  
 data = train\_set)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

confusionMatrix(predict(model2, test\_set), test\_set$Group)$overall["Accuracy"]

## Accuracy   
## 0.88

# Indeed, more variables improve the model.

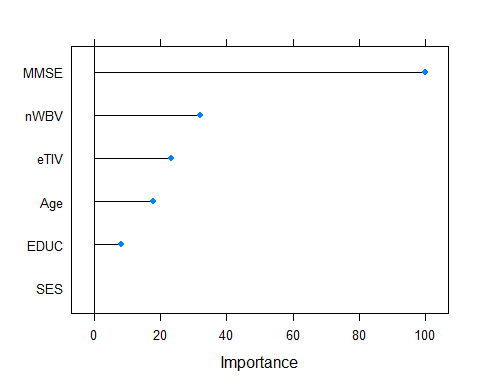
# Other algorith are tested, for example Random Forest. We get a better accuracy than just MMSE model and glm.

set.seed(1, sample.kind="Rounding")

model3 <- train(Group ~ Age + EDUC + MMSE + SES + eTIV + nWBV,  
 method = "rf",  
 data = train\_set)  
  
confusionMatrix(predict(model3, test\_set), test\_set$Group)$overall["Accuracy"]

## Accuracy   
## 0.8933333

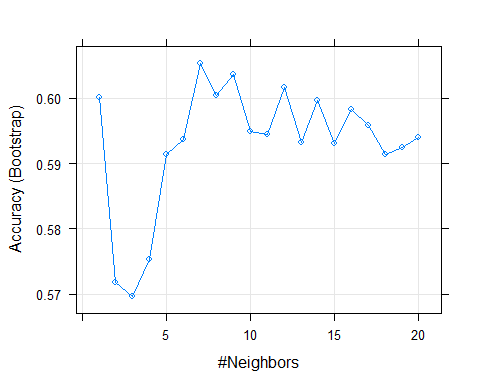
plot(varImp(model3))



# Another algorithm is KNN but the accuracy is only 0.627 with k = 7 as best “k”.

set.seed(1, sample.kind="Rounding")

model4 <- train(Group ~ MR.Delay + Age + EDUC + MMSE + SES + eTIV + nWBV,  
 method = "knn",  
 tuneGrid = data.frame(k = seq(1, 20, 1)),  
 data = train\_set)  
plot(model4)



confusionMatrix(predict(model4, test\_set), test\_set$Group)$overall["Accuracy"]

## Accuracy   
## 0.6266667