Assignment 4

Megan Marziali

## Part 1: Implementing a Simple Prediction Pipeline

### Loading required paackages

The following code chunk loads the required packages for the assignment.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.0 ✓ purrr 0.3.4  
## ✓ tibble 3.0.1 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.0 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(Amelia)

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.6, built: 2019-11-24)  
## ## Copyright (C) 2005-2021 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

### Loading data into environment and cleaning

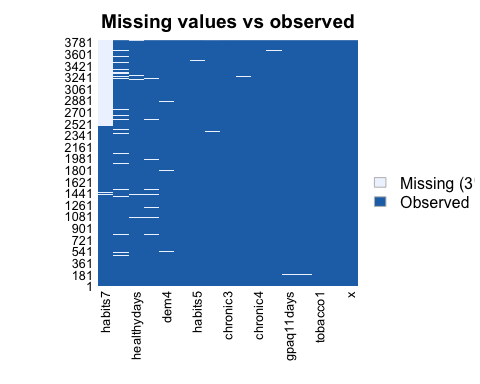
nyc\_data =   
 read.csv("./data/class4\_p1.csv", na = c("", ".", "NA")) %>%   
 janitor::clean\_names() %>%  
 mutate(  
 bmi = as.numeric(bmi),  
 gpaq8totmin = as.numeric(gpaq8totmin),  
 gpaq11days = as.integer(gpaq11days),  
 healthydays = as.integer(healthydays)  
 )  
  
summary(nyc\_data)

## x chronic1 chronic3 chronic4   
## Min. : 1.0 Min. :1.00 Min. :1.000 Min. :1.000   
## 1st Qu.: 953.5 1st Qu.:1.00 1st Qu.:2.000 1st Qu.:2.000   
## Median :1906.0 Median :2.00 Median :2.000 Median :2.000   
## Mean :1906.0 Mean :1.71 Mean :1.888 Mean :1.938   
## 3rd Qu.:2858.5 3rd Qu.:2.00 3rd Qu.:2.000 3rd Qu.:2.000   
## Max. :3811.0 Max. :2.00 Max. :2.000 Max. :2.000   
## NA's :28 NA's :17 NA's :8   
## bmi tobacco1 alcohol1 gpaq8totmin   
## Min. : 9.09 Min. :1.000 Min. :1.000 Min. : 0.00   
## 1st Qu.:23.05 1st Qu.:3.000 1st Qu.:2.000 1st Qu.: 0.00   
## Median :26.47 Median :3.000 Median :3.000 Median : 0.00   
## Mean :27.32 Mean :2.736 Mean :2.642 Mean : 43.77   
## 3rd Qu.:30.23 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.: 60.00   
## Max. :99.83 Max. :3.000 Max. :3.000 Max. :960.00   
## NA's :73 NA's :3 NA's :17 NA's :5   
## gpaq11days habits5 habits7 agegroup   
## Min. :0.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:2.000   
## Median :5.000 Median :2.000 Median :3.000 Median :3.000   
## Mean :4.273 Mean :1.972 Mean :2.734 Mean :2.781   
## 3rd Qu.:7.000 3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :7.000 Max. :4.000 Max. :5.000 Max. :4.000   
## NA's :5 NA's :17 NA's :1335 NA's :8   
## dem3 dem4 dem8 povertygroup healthydays   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.0 Min. : 0.00   
## 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:2.0 1st Qu.:27.00   
## Median :2.000 Median :2.000 Median :1.000 Median :3.0 Median :30.00   
## Mean :1.594 Mean :1.777 Mean :1.386 Mean :3.3 Mean :26.34   
## 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:5.0 3rd Qu.:30.00   
## Max. :2.000 Max. :2.000 Max. :2.000 Max. :6.0 Max. :30.00   
## NA's :30 NA's :6 NA's :244 NA's :88

nrow(nyc\_data)

## [1] 3811

missmap(nyc\_data, main = "Missing values vs observed")



Based on missingness, I would not include the variables “habits7” and “povertygroup” in any models. I would also remove all missingness from variables that I intend to keep in models. With all variables and missing observations included, the total N for the dataset is 3811.

nyc\_restr =   
 select(nyc\_data, -habits7, -povertygroup) %>%   
 na.omit()  
  
nrow(nyc\_restr)

## [1] 3552

After removing the variables “habits7” and “povertygroup”, and deleting missing observations, the total N for the dataset is 3552.

### Data partitioning

set.seed(100)  
  
train.indices = createDataPartition(y = nyc\_restr$healthydays,p = 0.7,list = FALSE)  
  
training = nyc\_restr[train.indices,]  
testing = nyc\_restr[-train.indices,]  
  
#Do I need this bit?  
#training.2 = bc.data %>% dplyr::sample\_frac(.7)  
#testing.2 = dplyr::anti\_join(bc.data, training.2, by = 'id')

## Problem 1:

Fit two prediction models using different subsets of the features in the training data. Features can overlap in the two models, but the feature sets should not be exactly the same across models. Clearly state which features were used in the two models.

model.1 <- lm(healthydays ~ chronic1 + chronic3 + chronic4 + bmi + gpaq8totmin + gpaq11days + habits5 + dem3, data = training)  
summary(model.1)

##   
## Call:  
## lm(formula = healthydays ~ chronic1 + chronic3 + chronic4 + bmi +   
## gpaq8totmin + gpaq11days + habits5 + dem3, data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.6331 0.3866 2.0994 3.5958 11.3321   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.597303 2.023526 8.202 3.75e-16 \*\*\*  
## chronic1 2.169564 0.359629 6.033 1.85e-09 \*\*\*  
## chronic3 2.214862 0.519365 4.265 2.08e-05 \*\*\*  
## chronic4 2.988871 0.649312 4.603 4.37e-06 \*\*\*  
## bmi -0.050116 0.025593 -1.958 0.05032 .   
## gpaq8totmin 0.001162 0.001697 0.685 0.49339   
## gpaq11days 0.150963 0.058346 2.587 0.00973 \*\*   
## habits5 -1.295627 0.193862 -6.683 2.88e-11 \*\*\*  
## dem3 -0.451693 0.307560 -1.469 0.14206   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.481 on 2478 degrees of freedom  
## Multiple R-squared: 0.08678, Adjusted R-squared: 0.08383   
## F-statistic: 29.44 on 8 and 2478 DF, p-value: < 2.2e-16

model.2 <- lm(healthydays ~ tobacco1 + alcohol1 + bmi + gpaq8totmin + gpaq11days + habits5 + dem3, data = training)  
summary(model.2)

##   
## Call:  
## lm(formula = healthydays ~ tobacco1 + alcohol1 + bmi + gpaq8totmin +   
## gpaq11days + habits5 + dem3, data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.7803 0.3876 2.3832 3.8323 10.3809   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 32.9718396 1.2995072 25.373 < 2e-16 \*\*\*  
## tobacco1 0.4614780 0.2397705 1.925 0.05439 .   
## alcohol1 -0.8509997 0.2904958 -2.929 0.00343 \*\*   
## bmi -0.1078506 0.0252491 -4.271 2.02e-05 \*\*\*  
## gpaq8totmin 0.0005348 0.0017175 0.311 0.75551   
## gpaq11days 0.1913265 0.0591128 3.237 0.00123 \*\*   
## habits5 -1.4746880 0.1962857 -7.513 8.02e-14 \*\*\*  
## dem3 -0.3819569 0.3168549 -1.205 0.22814   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.607 on 2479 degrees of freedom  
## Multiple R-squared: 0.05533, Adjusted R-squared: 0.05266   
## F-statistic: 20.74 on 7 and 2479 DF, p-value: < 2.2e-16