HW 4, STAT 452
Due: Thursday, March 11

Reading: Chapter 4, pp. 127–137, from *An Introduction to Statistical Learning*

Directions: Please submit your completed assignment to Blackboard. The assignment should be completed using R Markdown and rendered to an HTML or PDF format.

Exercise 1. Suppose you build a classifier that predicts whether an email message is `spam` or `not spam`. The table below shows the actual and predicted values for 10 emails. Based on these results, calculate the accuracy, sensitivity, and specificity.

	Actual	Predicted
1	`spam`	`spam`
2	`not spam`	`not spam`
3	`not spam`	`spam`
4	`not spam`	`not spam`
5	`spam`	`spam`
6	`spam`	`not spam`
7	`not spam`	`not spam`
8	`spam`	`spam`
9	`not spam`	`spam`
10	`not spam`	`not spam`

Exercise 2. For this exercise, use the `NHANES` data. First, run the following code as a pre-processing step to remove missing data and recode `PhysActive` (0=No, 1=Yes):

```r
library(tidyverse)
library(NHANES)
nhanes2 <- NHANES %>%
  select(PhysActive, BMI) %>%
  na.omit() %>%
  mutate(PhysActive = ifelse(PhysActive == "Yes", 1, 0))
```

(a) Fit a simple logistic regression with `PhysActive` as the binary response variable and `BMI` as the predictor. Use `summary()` to print the results, and write down the equation for the estimated logistic regression model.

(b) Use `ggplot2` to plot the estimated logistic regression curve for the probability of being physically active as a function of BMI.

(c) What is the predicted probability that a person with a `BMI=20` is physically active? What is the predicted probability that a person with a `BMI=30` is physically active?

Exercise 3. For this exercise, use the 2016 election data for US counties:

```
county_votes <- readRDS(url("https://ericwfox.github.io/data/county_votes16.rds"))
```

(a) Randomly split the `county_votes` data frame into a 70% training and 30% test set. Make sure to use `set.seed()` so that your results are reproducible.

(b) Fit the following logistic regression model, which uses 8 demographic predictor variables, on the training set:

```
trump_win ~ pct_pop65 + pct_black + pct_white + pct_hispanic
                + pct_asian + highschool + bachelors + income
```

(c) Make a confusion matrix between the actual and predicted values on the test set. Then use the confusion matrix to compute the accuracy (percent correctly classified), sensitivity (percent of Trump wins (1) correctly classified), and specificity (percent of Trump losses (0) correctly classified). Use a 0.5 probability threshold when classifying each point (county) in the test set as a Trump win or loss.

(d) Plot the ROC curve and compute the AUC.

(e) In terms of the cross-validation results how does the multiple logistic regression model, which uses the 8 demographic predictor variables, compare with the simple logistic regression model from lecture 10, which uses `obama_pctvotes` as a predictor?

Bonus. [2 points] For this exercise, use the `NHANES` data. First, run the following code as a pre-processing step to remove missing data and recode `PhysActive` (0=No, 1=Yes).

```
nhanes3 <- NHANES %>%
  select(PhysActive, BMI, Age) %>%
  na.omit() %>%
  mutate(PhysActive = ifelse(PhysActive == "Yes", 1, 0))
```

(a) Randomly split the `nhanes3` data frame into a 70% training and 30% test set.

(b) Using the training set, fit a logistic regression model for `PhysActive` with `BMI` and `Age` as predictors.

(c) Make a confusion matrix using the test set data, and then compute the accuracy, sensitivity, and specificity.

(d) Plot the ROC curve and compute the AUC.

(e) How does the accuracy of the logistic regression model, with `BMI` and `Age` as predictors, compare with the null model, which would predict that every person is physically active?