Introduction to Data Science Homework 3

This question will investigate a simple classification problem with one parameter, and various aspects of confusion matrix and ROC curve. The data we will use are heights of men and women. (Again, this data set is not real, but generated; the code used is provided.)

We will guess a person's gender based on their height alone by the following model:

$$y = \begin{cases} \text{M } (+), & \text{if } x < x_0 \\ \text{W } (-), & \text{if } x \geq x_0 \end{cases}$$

1. To begin with, plot histograms to visualize height distributions for men and women.

2. Fit a logistic regression model to this data set. What is the value of height x_0 that gives $p(x_0) = 0.5$? (This is the best threshold value to guess a gender of a person based on height alone.)

   Hint 1: Calculation of the likelihood function might be a little complicated. See lecture notes.
   Hint 2: To avoid underflowing, it is a good idea to work in term of logarithm of likelihood instead.
   Hint 3: Also keep in mind that when a quantity is maximized, its negative is minimized.

3. We will now vary the threshold value, and evaluate various confusion-matrix-related quantities. Vary x_0 in fine grid, from its lowest to highest possible values (these corresponds to predicting everyone to be men and women respectively). For each value of x_0, calculate these parameters: TP, TN, FP, FN, accuracy, precision, sensitivity, specificity, F1-score. Finally, plot them as a function of x_0.

4. Two common plots used to evaluate the model performance are the "Precision-Recall curve" and the "ROC curve". Make these plots and find their areas under curve.