BUS/CSC 386 Homework #4

For HW #4, please download the Higgs dataset from Box using this link:

https://berea.box.com/s/plddx4e3vsmy6swd84lcu9k9gs74om3q

This data is from the Large Hadron Collider (Atlas Detector), with label as a binary response variable. An "s" indicates the presence of tau-tau decay possibly indicating a Higgs boson while "b" indicates background noise. All other columns are predictors except EventID.

1. Import the Higgs.csv data set into RStudio using the read.csv function. Open a Word document, put every team member's name on it, and save it as Team # Homework_4 where # is your team number.
2. Examine the dataset for number of samples, variable types, number of predictors, and type of response variable. Put that information in your Word document.
3. Check the dataset for missing values and determine a strategy should they exist. Put that information in your Word document.
4. Split the dataset into randomized training and test sets, with 80% for training and the remainder for test.
5. Bake the data to center and scale while keeping the response variable unchanged. Add any additional transformations that may be necessary to prepare the data for training.
6. Save the training and testing response variables into separate vectors and delete them from their respective datasets.
7. Develop four different neural networks, ranging from one to four layers (the first network will have one layer, the second two layers, etc.) and all other hyperparameters of your choosing. Reserve 20% of the data for validation in the fit command. Run each network for 20 epochs using the warm restart method every 5 epochs.
8. Describe each of the four different networks in your Word document and paste the loss/accuracy chart with those descriptions. Indicate issues with overfitting or underfitting. After training the four networks, choose the most accurate model and run it against the testing data. Enter the final test accuracy into your Word document.
9. Summarize your report with ideas why the most accurate network you developed was the most accurate; in other words, try to understand how the network learned the data.