BUS/CSC 386 Homework #3

For HW #3, please download the supersym dataset from Box using this link:

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

This data is from a physics experiment, with column #1, target, as a binary response variable. One indicates the presence of supersymmetry while zero indicates its absence. All other columns are predictors.

1. Import the supersym.csv data set into RStudio using the read.csv function. Note: this data is already randomized. Open a Word document, put every team member's name on it, and save it as Team # Homework_3 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. 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.
5. Save the response variable into a separate vector and delete it from the dataset.
6. Split the baked data set and associated response vector into four similar data sets with 200-250k samples each. Note: you don't have to create a test dataset, only four equal sized training sets.
7. Develop four different neural networks and test each individual data set using a different network. 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.
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.