MACHINE LEARNING

There are now five rules that apply to all projects so far:

a) Follow instructions *precisely*. If I do not tell you what to write on a particular line, leave it blank.
b) All code must be *scalable by sample size* unless specifically noted otherwise.
c) Any code using *magrittr* should contain a max of one verb per line.
d) Always use standard headings unless otherwise specified: **R Studio API Code, Libraries, Data Import and Cleaning, Analysis, Visualization.** Delete any headings without code under them.

Part 1 – Data Acquisition and Preparation

1. Download the 2006 General Social Survey from http://gss.norc.org/Documents/spss/2006_spss.zip and place the SPSS data file this archive contains in your project. You will likely also need the codebook from http://gss.norc.org/documents/codebook/gss_codebook.pdf
2. In **week11.Rmd**, clean the survey appropriately, which will include at a minimum:
    a. Converting from the SPSS format *in R* (do not use SPSS)
    b. Ensuring any missing, don't know, inapplicable, or otherwise not-clearly-answered items are marked as missing according to R
    c. Converting any variables you will need to use into appropriate data types or other formats
    d. Isolating the variables we will be using as your final dataset:
        i. the personality inventory
        ii. the respondent's self-reported health
    e. Annotating all the actions you took and why you did them the way you did.

Part 2 – Machine Learning

1. Create a 250-case holdout sample. Ensure you do not include these cases in model development.
2. Run ordinary least squares regression to predict health from all personality variables and all of their 2-way interactions. Calculate 10-fold cross-validation statistics and also prediction in holdout.
3. Run this same model (health on personality) three more times, but using 10-fold elastic net regression, support vector regression, and extreme gradient boosted regression. Report cross-validation and holdout performance. How did your results change between models? Why do you expect this happened, specifically? Explain which tuning parameters worked best for the elastic net model (i.e., what did they mean?).
4. Create summary tables and figures comparing the success of all four machine learning approaches on appropriate metrics. Ensure your results are in fact directly comparable.
5. Of the four, which model do you prefer, and why? Are there tradeoffs? Write at least one paragraph.
6. For this section, here are some hints:
    a. You can treat the original health codes (1-4 or 2-5) as your continuous DV for this project.
    b. You will need to add additional pre-processing to do run this analysis, but this may or may not be done in *caret*.
    c. You should probably Google and/or search caret docs for algorithms you don't know.
    d. Do not listwise delete. Consider missingness explicitly and handle it appropriately.
    e. Explain what you've done and why you made each decision you made in your notebook. Include justifications and interpretations of all results and respond to all questions.