HW7 [Your Team Name Here]

[Your Names Here]

Due April 14, 2017, late acceptance until April 15

We will return to binary regression with the National Election Study data from Gelman & Hill (GH). (See Chapter 4.7 for descriptions of some of the variables and HW3 for initial model fitting).

[*The following code will read in the data. Remove this text and modify the code chunk options so that the code does not appear in the output.*]

```
# Data are at http://www.stat.columbia.edu/~gelman/arm/examples/nes

nes <- read.dta("nes5200_processed_voters_realideo.dta",
                    convert.factors=F)
#  Clean data

# filter data to include year, age, income, race, white, black, female,
# religion, south, state, region, maritial_status, party afilliation
# (as in HW3) ideology (as in HW3).
# The response should be 0/1  with 1 = vote republican in the
# presidential election.

#  remove NA's

# convert variables that are coded as numerical as
# factors (state, region, etc)

# create random split 50% sample for test and training
set.seed(42)

# Note the variable state has more than 50 levels, and I
# have not heard back from authors about a data dictionary.
# You may decide how to handle this; ie. assume 1:50 are US and
# others are territories other locations and use the 1:50 ???
# just document what you do.  (they are not sorted alphabetical)
# Discuss how this limits your modelling
```

1. Using the the training data, fit a tree model to the data to predict probability of voting republican in the election and prune. Comment on the selected tree - which variables are important? are there interesting interactions or clusters? (provide graphics or tables to highlight findings)
2. Using the the training data, fit a random forest model to the data to predict probability of voting republican in the election. Comment on the results - which variables are important? what insights does the model provide (support with graphics if possible)?
3. Repeat 3, but using boosting.
4. Repeat 3, but using bart. Comment on any partial dependence plots or other output that is of interest in explaining the model.
5. Using `gam` or `bam` from `mgcv` fit a generalized additive model to predict probability of voting republican using smoothing splines for fitting examining nonlinear functions of the continuous variables. Are there any interactions that you might expect will be important (based on tree models or information from Ch 14 Gelman & Hill?) In `mgcv` you may allow different curves for levels of a factor using the `by`

option: `race + s(age, by=race)`. Random intercepts for say state, may be obtained via `s(state, bs="re")`. Using residuals, residual deviance, AIC, or other options find a predictive model that seems to be reasonable for the training data, exploring non-linearity, random intercepts and slopes. Provide a brief description of how you came up with your final model and describe what insights about voting it provides.

6. (optional) Using any insights from the models, fit a model in JAGS that seems to capture the best features of the models above or addresses any deficiencies that you see.

7. Using the models from 1-7 (8), determine the error rate for each model for predicting on the test data.

8. Provide a summary of your findings. Your comments should address benefits and advantages for the different methods. Which method has the best predictive accuracy? Which provides the most interpretability or insight into quantifying factors? In explaining your findings and insights provide graphs and tables that help quantify uncertainty and illustrate effects of the different characteristics. (Using the training data and any of the models above do you reach similar conclusions as in Ch 14 of Gelman and Hill?)