Marc Alessi
ATS 780

Due: 9/18/23
HW1 Checkpoint

Goal: Predict whether a U.S. Congressional District will vote Democrat (DEM) or Republican (REP; baseline: 2019 Cook Political Voting Index).

Github link: https://github.com/mja244/ats780/tree/main

Tasks completed so far:

- Compiled data
  - Main set of predictor data from the 2012, 2014, 2016, 2018, and 2020 U.S. Census estimates. Features include data on gender, education, race, ethnicity, and language.
  - 3 extra features not from U.S. Census estimates:
    - U.S. Generic Congressional polling average before each House election from Real Clear Politics
    - Individual district polls from the New York Times, Siena College, or Emerson (only available for a few districts in 2018 and 2020)
    - 2019 Cook Political Voting Index
  - In the beginning, I separated out 2020 data to use as testing data.
- Prepped data
  - See file "data_prep.py" in HW1 folder
  - Split remaining elections (2012, 2014, 2016, 2018) into training/validation data
- Started to train a Random Forest model
  - Originally, I started with a Random Forest Regressor: I set REP district results equal to -1 and DEM district results equal to 1. The model would predict a value between -1 and 1, and I would set the district to REP or DEM depending on if it was positive (DEM) or negative (REP). The mean absolute error of this model was greater than the baseline.
  - I switched to a binary Random Forest Classifier (DEM or REP as the choices), and this seems to work much better.
- Parameter testing
  - Still experimenting with parameters overall
    - Lowest values in SW and NE corners of the confusion matrix when tree number is set to 16 (when testing on validation data).

Issues I'm having with the model:

- My model is not doing well with predicting how swing districts will vote. This appears to be a problem with overfitting with the training data.
  - However, the model is doing quite well in general. With an ideal number of trees (12) in the RF, I get a precision of ~.90 and a recall of ~.91. But these high precision/recall scores are not necessarily good. Predicting how a swing district votes in an election is the most important because it decides who holds the House.

- I also have not split up my training data into training and validation data, and I have mistakenly been using the testing data as what I should be using for validation data (with tuning parameters and whatnot).