STAT 195 Project Part 1, Fall 2018

This class has no exams, and much of the course grade will instead be determined by a two-part project (this assignment is the first part of that project). The goal of the project is to put the theoretical knowledge and tools taught in this class into practice, and also to get students thinking beyond the material taught in class, as the subject of statistical machine learning has many facets we will not have time to cover in sufficient depth in class. Since this class is for upperclassmen, students should be capable of teaching themselves to a certain extent when they come across problems beyond the scope of the class, but everyone is *strongly* encouraged to take advantage of the TFs' and instructor's office hours to ask about and talk through such problems. This project will also rely heavily on the computational skills developed in the prerequisite class, CS 181.

Due: Components of this project part will be due October 5th, October 12, October 19, and November 5, all at 11:59 PM on Canvas (no late days allowed)

Groups: You may work on this project in groups of no more than three, but each group should turn in just one submission of each component.

Background: The United States midterm elections will be held November 6, 2018, and will decide all 435 voting seats in the House of Representatives, 35 of the 100 seats in Senate, and many statewide and local government positions as well, including most Governorships. There were some special elections for the House of Representatives held before November 6 which this project is not about, as all those seats will still be up for election on November 6. Also note that the six seats in Louisiana are special because each of its November 6 elections do not actually elect a representative unless a candidate wins over 50% of the vote. This assignment will focus on forecasting/predicting just the House of Representatives, and a wealth of information is available online about the elections and what information might be predictive of it. Here are some good resources to get you started:

- Wikipedia has a comprehensive list of basic information, including which candidates are running, their party affiliations, who the incumbent is, etc.
- FiveThirtyEight does a forecast of the House of Representatives, and there are quite a lot of articles and podcasts you can read/listen to in order to learn what Nate Silver thinks is a good idea to do.

Assignment: The general assignment is to forecast the 435 seats in the House of Representatives which are up for election on November 6, 2018. Note that for Louisiana's six seats, one of the options will be that there is no winner on November 6. The assignment will have three parts:

- *Feature downloading* (Feature choice due Friday, October 5 at 11:59 PM; feature due October 12 at 11:59 PM): One of the challenges of building your own forecast is choosing and downloading data, and to make this less of a burden for everyone, each group will download one feature and then all these features will be shared with all the groups. To be explicit, each group will download a single unique feature of their choice with (a) at least four elections'-worth (8 years, since Congressional elections are held every two days) of training data, with at least 80% of elections in the training data containing feature values for at least the Democratic and Republican candidates (or just one if one party didn't have a candidate), and (b) the feature values for at least the Democratic and Republican candidates (or just one if one party doesn't have a candidate) in each of this year's 435 elections. The features for this year's candidates, if time-varying, should all be from one time point and should be current as of at least October 5.

  To ensure uniqueness, each group will first email the TFs with their chosen feature and an explanation for why that feature is relevant to the problem, and if more than one group has the same idea for a feature then precedence will be given to the group that emailed first and the other groups will need to think of another feature. The TFs will try to maintain a relatively up-to-date list of features that have already been claimed by groups so groups can avoid wasted work. Each group needs to have a unique feature chosen by October 5 at 11:59 PM, but the feature itself is not due until October 12 at 11:59 PM. The feature and its training data should be submitted as a CSV file with four columns: the election year, the district, the candidate name, and the feature value. The feature should also be submitted with a short explanation (one paragraph) of where the feature came from and how it was downloaded.

  I don't want to constrain what you think might be useful features, and as long as you can justify your feature to the TFs' satisfaction (and find sufficient training data), it is fine with me. And binary features are OK too, e.g., incumbency status. But here are two resources to get you started:

  - The Library of Congress maintains a page with links to election data going back many years.
  - The Harvard Kennedy School maintains a page with links to both free and Harvard-specific election data.

  Everyone is still highly encouraged to find and use other features beyond those that are shared with everyone, but at least this gives everyone a good place to start.

- *Early forecast* (due October 19 at 11:59 PM): You will produce three products, to be submitted on Canvas:

  - A matrix with 435 rows and three columns: the first is a list of seats of the House of Representatives, ordered first alphabetically by state and then numerically within state; the second column is your projected winner of that seat (for each of Louisiana's six seats, one option is "No Winner"); the third column is the probability you assign to that projected winner.

  - A pair of probabilities, one each for a Republican and a Democratic majority in the House of Representatives following the November 6 election (note these two probabilities need not sum to one).
  - A written report explaining the above two forecasts, not to exceed one page of single-spaced, 12pt text with one-inch margins on all sides (not including references, though references are not needed for this report). The report should briefly explain the choice of features and training data, and the choice of machine learning method. It should not contain any code or figures. You should justify as many of your choices as possible with domain knowledge but do not need to cite sources for that information. For any choices not made entirely based on domain knowledge, but made in part or wholly in a data-driven way, you should explain (briefly) how the data was used to make that choice.

- *Last-minute forecast* (due November 5 at 11:59 PM): You will produce anew the same two forecasts as for the early forecast, and an expanded and more detailed report with a page limit of 10 pages (not including references) that explains both the early and last-minute forecasts. This report should expand on the early forecast report and go into more detail, as well as give a detailed explanation of the last-minute forecast as well, including any changes in approach/choices between the early and last-minute forecasts. For this report, choices justified by domain knowledge should cite credible sources presenting that information, and data-driven choices should be explained fully (e.g., in the early forecast report, justifying the choice of a tuning parameter "by cross-validation" would be sufficient detail, while for the expanded report you would need to specify how many folds, the loss function, how any missing data was handled, etc.)

Grading: This assignment will be jointly worth 30% of the course grade, and all group members will receive the same grade. The grading out of 100 points will be divided as follows:

- Feature downloading: 10 points. This part is really to help focus this project more on the machine learning and less on data scraping for everyone. So as long as your feature meets the criteria detailed above and is well-justified and its procurement explained, you will likely receive 9-10 points.

- Early forecast: 30 points. You will not be explicitly graded on the performance of your forecast, but how much thought and effort went into the forecast, how well you apply concepts covered in class, and if/how you went beyond the course material to solve other issues that arose. For example, a forecast that simply copies FiveThirtyEight's would likely perform fairly well but merit a low grade.

- Last-minute forecast: 60 points. The same principles as for the early forecast apply, but with a substantially higher burden of justification. An ideal forecast/report would have detailed justification based on training data or domain knowledge for every choice made in the process of constructing the forecast, with all choices clearly and concisely explained, including high-quality sources and figures if necessary.