Dennis Sun | **DATA 401:** Data Science | Alexander Dekhtyar

Project 3
Classification

Due: Monday, December 12

1 Making Sense out of the Elections

Eight days after the presidential elections, and countless "what went wrong" articles that a variety of outlets have, no doubt, published, t is clear that predictive analytics for this particualr election suffered a *massive loss*.

Let us try to figure out what happened.

The basic steps are as follows. Consider the USA at the granularities of counties and/or congressional districts (you may get better view into what happened if you look at the county-level data).

Collect 2012 and 2016 US Presidential elections data at the selected level of granularity.

Collect demographic data (best available) for each county.

Collect any other county-level data you may be able to find, and you think may be useful.

Using the data collected, try to figure out what the predictive model predicting the winner of each county should have been.

Contrast the 2012 election with the 2016 election.

Compare your observations with what the others (there will be A LOT of post-election "where did we go wrong" analysis) are saying.

Using analytical techniques studied in class, organize the US counties into clusters based on their demographic and other features. Determine which counties voted "with their cluster" and which counties did not.

Collaboration.

For this project, we allow limited collaboration between the teams. The allowed collaboration is:

- Data and data acquisition sharing. Establish a Piazza thread, post links to discovered data and feel free to share any data acquisition code /code snippets. Teams that choose to use the data sets you built, or the acquisition code you developed must credit the source in their reports.
- Discussions. We encourage teams to discuss this part of the project.
- Ensembles. To a large degree, if the entire class can come up with one convincing model of what happened, it will be much better than seeing five weak models. The teams may attempt to combine the models they developed into various ensembles.

Notes. I recommend a burst of activity related to procuring the minimum necessary data to conduct this analysis ASAP. You can use preliminary results for the 2016 elections, as the final results won't be certified until a few weeks from now. Unless we are in a 2000-style recount territory (which this election does not appear to be heading for) the preliminary data should be sufficient as the prediction target.

Concentrate on using classification and clustering techniques. While most election projection models used regression and Monte Carlo simulations, we want to stress the ability of the demographic information about a county to predict the winner of the vote. Note that this is a *weaker* prediction, becasue simply predicting county winners does not allow one to predict the final Electoral College scores. If you are comfortable at your ability to predict county winners, you can try using regression methods from prior Projects to predict the margin of votes in each county. At the same time, clustering will allow you to observe groups of similar counties irrespective of their voting behavior and then to see how well the voting behaviors of specific clusters can be predicted.

We want models with *explanatory power*. That is, you should attempt to extract some explanations from the most accurate models you can build.

What it should look like. Traditionally, election predictions are based on turnout models: that is a predicted distribution of voter turnout broken into demographic categories for which the voting behavior is predictable with much better accuracy, combined with the specific predictions of how each demographic group will vote. In Election'2106 two competing models have emerged: the conventional widsom model to which most of the election analysts subscribed suggested the turnout model similar to the 2012 election (where, too, the eventual margin of victory for the winning candidate was

significantly underestimated), and the competing model that suggested a drastically different turnout model.

As seen from the election results, the competing model appears to have been more correct.

This does not mean however, that the second model provided proper predictions about the turnout *everywhere*. It appears that in fact, the population of the USA "fractured" in terms of turnout into the states (and counties) that followed the "expected" model, and states (and counties) that followed the "alternative" model.

We would like you to determine which geographic locations followed which model. You can use the 2012 election results as the proxy for the "expected" 2016 model, and attribute major deviations in voting behavior from the 2012 patterns to the "alternative" model.

Deliverables.

You have two types of deliverables: code and project reports.

Code deliverable is for archival purposes only. Submission instructions will be provided to you closer to the submission deadline.

Reports shall be text-processed and submitted in PDF format. The reports shall contain your team's narrative for each of the parts of the project. We want to know what you have tried, what worked and what did not. We also want to see, the most accurate predictions you were able to make highlighted in your report.

The reports can be written in an informal style (i.e., do not treat them as academic papers), but they should be as complete as possible. You should find ways to visualize important insight you gained and include these visualizations into the report.

The project presentations will take place during the final examination period reserved for this class (*Monday, December 12, 1:10-4pm*). As such, each team will be given about 25 minutes to present the findings. This is more time than the presentations for the other projects in the class.

Each team must prepare and deliver a *professional presentation* using standard presentation software[1] The presentation shall contain all materials the team wants to present about its work. Try to find a reasonable part of the presentation for each member of the team to deliver.

Your presentation may include demos of code (Jupyter notebooks, for example), but those demos must be fast and validated. Remember - you are delivering a professional presentation.

[1]For example, MS Powerpoint, Keynote, or Google Slides.