Decision Desk HQ Fellow Data Exercise

Introduction

Here at Decision Desk HQ (DDHQ) we spend a lot of time thinking about elections: who is likely to win them, when will we know, and where will all the data come from. During the 2018, 2020, and 2022 election cycles, one of our biggest projects was our seat-by-seat public forecast for congressional control, and - in 2020 - for the presidential race. For example, you can see our 2022 congressional forecast here.

Project

Pretend it is September of 2018. We've collected a large quantity of historical information about US House races stretching back to 2006. We need to predict the margin-of-victory for each of the 435 congressional seats composing the House of Representatives.

We have attached a file - DDHQ_Data_Exercise.csv - containing all of this historical information from 2006 to 2018. For the races from 2006 to 2016, we've also included the final result. These numbers are not fake - this is all real data for these districts across this time period. There are no tricks here.

For the rows corresponding to 2018, the Margin-of-Victory column is filled with a "???." Your job is to fill this column in with your prediction for each House district. A data dictionary for the dataset we're providing is given below.

Specifically, we want you to provide three different items in your final submission:

1. The file DDHQ_Data_Exercise.csv, in precisely the same form we've provided, but with your predicted Republican-Minus-Democratic victory margins replacing the "???" icons in column F.
2. A report where you characterize your methodology and process for obtaining your predictions, alongside an interpretation of your findings. Your report will be read by a team of data scientists and elections experts, so you can assume that your readers are familiar with machine learning terminology and American politics. That said, please provide any information that might help us understand what you did. Please describe what kind of model you built, why you choose that kind of model, and how you produced your predictions. A picture is often worth a thousand words.
3. Whatever code you used to perform your analysis. We are not picky about format. People often provide either a set of python scripts, a set of R scripts, or a Jupyter notebook, but these specific coding languages are not required. You are absolutely allowed to create new engineered features, but please note this in your code.

Table 1: Data Dictionary

Variable	Description
Race ID	Unique identifier for each House race in each year.
Chamber	Chamber of Congress. For this exercise, this will always be the US House.
State	State where race occurs.
Congressional District	Congressional district number for race
Year	Election cycle year
R-D Victory Margin	Republican vote share minus Democratic vote share. A positive number here means the Republican won.
Incumbent Running?	Indicates if an incumbent Republican or Democrat ran in the elec tion, or if seat was open.
Geography	Description of district's urban/rural composition.
District PVI	The Partisan Voter Index for that district in that year.
Population Per Square Mile	District population density
Share African-American	Share of district population that is African-American
Share Asian	Share of district population that is Asian
Share College Educated	Share of district population with college degree
Share Hispanic	Share of district population that is Hispanic
Share Non-Hispanic White	Share of district population that is White
Unopposed D?	Indicates if a Democrat ran unopposed for House in that district that year.
Unopposed R?	Indicates if a Republican ran unopposed for House in that district that year.
Vote Share of Last D Presidential Candidate	Share of the Vote that the last Democratic presidential candidate received in the district.
Vote Share of Last R House Candidate	Share of the Vote that the last Republican House candidate received in the district.

Q & A

1. **How long do I get to complete this project?**
   As long as you need! After you take a look at the data, send us an email telling us how much time you think you'll need to complete the project. People often ask for 1-2 weeks, but there is neither reward for going faster, nor penalty for going slower.

2. **Am I only allowed to use the data given in the input file to generate my model and predictions?**
   No! You are welcome to incorporate any additional features or data you find into your model, if you think it's appropriate (that might be polling, more demographic data, economic data...whatever). Please just explain what you did in your report, and cite where you got any outside data from.

   Remember that you are building your model in September 2018, so please only use data you might have had at that time. For example, you would not have known anything about election results from 2020 in 2018.

3. **Are there any "gotchas" in this assignment?**
   Only the ones provided by real life! None of the data here is intentionally manipulated or altered. We have not intentionally inserted any errors for you to "catch." This is all real data going back to 2006. You are using real data to make a real forecast for the 2018 midterm election.

4. **Am I allowed to check my work?**
   I definitely can't stop you! You can easily google the final results of the 2018 US House elections and see how your model is performing as a check. We do not expect your model to accurately predict the result in every House race. Our team is composed of multiple election experts, and we know which races are hardest to predict correctly.

5. **How is this graded?**
   This is not a pass/fail test. Our team will look at all of your material, and we will determine next steps holistically. Some things we will take into account: Code quality, written communication skills, methodological approach, and model accuracy. Our assessment takes all of these factors into account.

6. **If I get stuck, am I allowed to ask questions?**
   Absolutely! If you have any questions or concerns, please email Kiel Williams (kiel@decisiondeskhq.com) and Patrick McCaul (patrick@decisiondeskhq.com), and we will be happy to provide some guidance.