Øptimus Data Science Predictive Modeling Exercise

INTRODUCTION

Here at Øptimus, we spend a lot of time working with voter files, or datasets consisting of information about registered voters. (You can find more here: http://en.wikipedia.org/wiki/Voter_file) We've included a pared down voter file for you to work with (voterfile.csv) consisting of 50,000 records from Nevada.

PROJECT

Now that you have the data, here is the problem: Pretend it is May 2014. We need to predict voter turnout forthe 2014 general election, held in November.

We would ultimately like to see four things from you:

1. Your predictions in .csv format (more info below).
2. A document where you interpret your results and specify the prediction methods you used.
    - This will beread by members of the data science team, so you can assume that we are comfortable with terminology. However, please include whatever else you think will aid us in understanding your prediction method, including why you chose it and how you implemented it. Your ability to communicate technical knowledge well is important as we may soon be working as part of a team!
3. A brief set of slides (no more than one cover slide and five substantive slides) where you present your results.
    - This should be geared toward members of a campaign team. Use your judgment to Communicate the most important points to people who do not have a statistics background.
4. The code that you wrote to generate your analysis.
    - This code should include everything that you do – reading the data in, using libraries, youranalysis, any checks, renaming variables, etc.

METHOD

Use whatever prediction method you feel is appropriate, but please specify the models or algorithms that you use. Though not required, we prefer that your models or algorithms be written in a scripting language such as R, Python, or Matlab. Finally, if you manipulate the original data, we give you or create new variables, please put a comment before the code that implements these changes.

.CSV RETURN FILE

The .csv file should contain the optimus_id variable plus all independent variables used to predict turnout. It should contain a field labeled vote with 0 if the person is predicted to stay home or 1 if they arepredicted to turnout. Finally, please include a field labeled vote_prob, which contains the estimated probability of turning out for each individual from 0.00 to 1.00. For example, it might look somewhat like the following:

optimus_id	age	vh14p	vh12g	vote	vote_prob
861681	69	1	0	1	0.729419
108469	20	0	1	1	0.635286
644435	28	0	0	0	0.238256
576830	78	0	1	1	0.604895
167371	68	1	0	1	0.752746
974034	69	0	0	0	0.164904
660415	53	1	1	1	0.768196

Øptimus Data Science Predictive Modeling Exercise

OVERVIEW OF VOTER FILE VARIABLES

Field	Description
optimus_id	Unique id assigned to each person
age	Age of registered voter
party	Registered political party
ethnicity	Modeled ethnicity
marital	Modeled marital status
dwellingtype	Dwelling type
education	Modeled education (commercial data)
cd	Congressional district (geography)
dma	Designated market area (geography)
occupationindustry	Modeled occupational industry (commercial data)
net_worth	Net worth (commercial data)
intrst_nascar_in_hh	Individual interested in NASCAR in household (commercial data)
petowner_dog	Likely to own a dog (commercial data)
intrst_musical_instruments_in_hh	Individual with musical interests in household (commercial data)
donates_to_liberal_causes	Donates to liberal causes (commercial data)
home_owner_or_renter	Homeowner or renter (commercial data)

Voter History

Field prefixed with vh contain vote history information where the two digits following vh refer to the year ofthe election, and the final letter p or g indicates whether the election is a primary or general election, respectively. For example, vh12g represents vote history information for the 2012 general election. A value of
1 should indicate that the individual showed up to the 2012 general election and a value of 0 should indicatethat the individual did not show up to the 2012 general election.

Historical Election Returns

The remaining fields refer to historical general (g) or primary (p) precinct-level turnout numbers.

Øptimus Data Science Predictive Modeling Exercise

Q & A

Q: How long do I have to complete this?

A: We realize you may have midterms, work, or other engagements, so we want to give you a reasonable timeframe for completing this project. **Please submit what you believe to be a reasonable timeline and your final project to kiel@0ptimus.com.** Should anything arise and you need a bit more time, please let us know.

Q: How long should I spend on this project?

A: That will depend entirely on you. Those of you who have experience in statistics may find the analysis easier,those of you who have experience coding may feel more comfortable working in a software package, and thoseof you who consider yourselves to be effective communicators may more quickly put together the write-up andpresentation. Some of you may take five hours, some of you may take a hundred hours. (Just kidding. Please don't spend 100 hours on this project. This is not meant to be that kind of project.) What I can say is that successful applicants will probably start this well before the deadline.

Q: How are we graded on this? Is it pass/fail?

A: This is not a pass/fail test. It is one of many factors we consider when selecting data science interns and fellows. We send the same exercise to both fellow and intern applicants, with obviously different expectationsfor both groups, but we also have your resumes and potential interviews with you. We grade each prediction exercise in terms of several facets:

- Accuracy (several different metrics for accuracy)
- Techniques/Methods
- Quality of Code
- Communication Ability

This is not an exhaustive list – we may consider other factors as appropriate – but this should give you a good idea as to what we are looking for in a candidate. We are looking for people who think logically, solve problemscreatively, and communicate effectively. How well you do is up to you!

Q: If I completely bomb this, do I still have a chance at Øptimus?

A: If you're applying for a fellowship, probably not. However, if you're applying for an internship, you absolutely still have a chance if you are not able to successfully complete this task. Like you've already read, thisis not pass/fail.

Q: So, what am I supposed to do again?

A: We want you to predict voter turnout for the **2014 general election**. In the voter file, this would be saved as vh14g (see above for explanation of voter file variables). Obviously, as it is May 2014, we do not have data forNovember 2014 and thus we cannot provide you with vh14g. We will leave it up to you as to how you choose to handle this.

Q: If I'm stuck, can I ask you questions?

A: Absolutely! The more you're able to complete on your own, the better, but we work in a team environment. Feel free to reach out to **kiel@0ptimus.com** if you have questions, need clarification, or believe there to be an error.