

**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 be read 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.

1. The code that you wrote to generate your analysis.

* This code should include everything that you do – reading the data in, using libraries, your analysis, 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.