Statistics 149 — Spring 2017 — Course Project

Mark E. Glickman

Key information:

Decision to work alone or as a team: 5pm on Friday, March 31, 2017

Prediction contest: Ends 10:00pm on Sunday, April 30, 2017

Written report: Due 10:00pm on Wednesday, May 3, 2017

Prediction contest web site: `https://inclass.kaggle.com/c/who-voted`

General description:

The final project has two components. The first involves your developing a predictive model on a data set I have made available on the web site *https://inclass.kaggle.com/*. The second part of the project involves your writing a short (no more than 6 pages of text) summary describing how you approached analyzing the data, how you converged on the final method you used to make predictions, and the substantive conclusions of your final model. Projects are to be carried out either on your own, or in groups of up to four students.

To make the prediction component of the project a bit more interesting, the *inclass.kaggle.com* web site allows you to post your predictions and see where you stand relative to others in the class on an ongoing basis as you continue to refine your predictive models. The individual or team that ends up with the most accurate predictions will receive an **all-expense paid dinner** with your instructor Mark Glickman along with the three TFs to a fairly modest restaurant during finals period (probably the Border Cafe). Maybe it's not as impressive as the $100,000+ prize funds connected with other kaggle contests, but it's better than nothing.

Initial steps:

You will need to set up an account on the *inclass.kaggle.com* web site. As long as your account has *harvard.edu* as part of the e-mail address connected to your login on the site, you should be able to join the competition without any special permission. If you cannot use your Harvard e-mail address or have problems accessing the competition, send me an e-mail and I can send you a link that invites you to join the contest. You may want to bookmark the competition URL for the duration of the project (see the URL in the "key information" above).

If you want to carry out the project with other people, you should begin the process of identifying with whom you want to work. I would like this process to be completed at latest by Friday, March 31. Once you have identified a teammate or teammates, you can add their logins from the contest dashboard (left column on the main contest page) by clicking "My Team." When letting me know your team composition, please also let me know the name of your team displayed on the kaggle leaderboard.

Prediction exercise description:

The goal of this project is to use the modeling methods you learned in the course (and possibly other related methods) to analyze a data set on whether a Colorado voting-eligible citizen ended up actually

voting in the 2016 election. These data were kindly provided by `moveon.org`. By clicking on the "Get the Data" link on the contest page, you will be able to download two files. The first, *train.csv*, contains a randomly selected 118,529 observations (75%) from the data set of 158,039 Colorado voting-registered citizens with the following variables:

1. `voted` (response variable: did the citizen vote [N/Y])
2. `gender` (gender [F/M], or unknown [U])
3. `cd` (congressional district)
4. `hd` (state house district)
5. `age` (age [years])
6. `dbdistance` (distance [miles] to nearest ballot dropoff location)
7. `vccdistance` (distance [miles] to voter's polling place)
8. `party` (D=Democrat, R=Republican, L=Libertarian, G=Green, O=American Constitutional Party, U=Unaffiliated)
9. `racename` (Race or religious affiliation)
10. `hsonly` (score for likelihood of having high school as highest completed degree)
11. `mrrg` (score for likelihood of being married)
12. `chldprsnt` (score for likelihood of having children at home)
13. `cath` (score for likelihood of being Catholic)
14. `evang` (score for likelihood of being Evangelical)
15. `nonchrst` (score for likelihood of being non-Christian)
16. `otherchrst` (score for likelihood of being another form of Christian)
17. `days.since.reg` (number of days since registered as a voter)

The score variables (`hsonly` through `otherchrst`), which take on values between 1 and 100, were derived from proprietary models based on phone surveys and information from public voter files. Higher scores indicate greater likelihood.

The second file, *test.csv*, contains the remaining 39,510 observations (25% of the original data set) with the same variables as above but with the variable `voted` withheld, and with the variable `Id` added. It is worth mentioning that some of the variables contain substantial missing data, including `dbdistance`, `vccdistance`, and the rating variables. You may want to investigate strategies for addressing the missing values.

Your job is to apply the model you developed on *train.csv* to obtain probability predictions on the withheld `voted` variable in *test.csv* as accurately as possible. The evaluation measure is described below. When you have determined a set of predictions, you should upload a .csv file containing your 39,510 probability prediction values in the same order as the observations in *test.csv* keeping the `Id` variable as part of the file. An example submission file is available on the data page. You can submit your prediction file from

the "Make a submission" link off the contest home page. You will then be shown the evaluation of your predictions using the evaluation formula based on a random 50% subset of the observations in *test.csv* (the same subset used for everyone), and your score will be placed on the leaderboard so you can compare your accuracy against others. Keep in mind that you can upload multiple prediction files throughout the contest period; your only limit is that at most five prediction files can be uploaded per day. So you have plenty of opportunities to improve your model predictions if others appear to be outperforming you. The *inclass.kaggle.com* site treats 5am ET as the day division – this translates to midnight UTC, so the five uploads per day can be made from 5am to 24 hours later. It is worth mentioning that because the scores reported on the leaderboard are based on only 50% of the test data set, the final accuracy (and leaderboard order) is likely to be a little different than the information posted while the contest is ongoing.

You should feel free to apply domain knowledge to the problem. For example, maybe certain interactions among variables are worth considering; perhaps female Democrats were more inspired to vote in Colorado than female Republicans.

Also keep in mind that Border Cafe fajitas are exceptional – this alone should inspire you to outperform the competition.

Prediction discrepancy evaluation measure:

Each .csv file you upload will contain 39,510 predictions. Let $\hat{p}_i$ denote the probability prediction for the i-th respondent in the test data set. The formula that will be used to compute prediction discrepancy is as follows. For each observation i, $i = 1, \cdots, n$, $(n = 39510)$ in the test data set, the total discrepancy is computed as

$$d = -\frac{1}{n} \sum_{i=1}^{n} \left(y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \right)$$

where y_i is the true (but withheld) indicator of whether potential voter i in the test data set actually voted. Your goal is to produce predictions that minimize d. Any real value of $\hat{p}_i$ between 0 and 1 is an acceptable prediction for each i.

The discrepancy measure d is larger when $\hat{p}_i$ and y_i are farther apart. On the leaderboard, you will see the discrepancies computed based on the 50% sample, or $n = 19755$ respondents. In addition to comparing your results to those of others in the class, I will upload "benchmark" predictions using simple predictive schemes.

Instructions for written summary:

The prediction exercise is to be accompanied by a written summary, which is due 3 days after the completion of the contest. You are advised to work on the written summary as you work on the prediction modeling exercise. The main goal of the written summary is to explain the final candidate models you used, the logic you followed that led to your final models, a brief description of some of the attempts and failures (if any) that resulted in changes in strategy, and substantive conclusions you learned about the probability of voting derived from your final model. The summary text should be no more than 6 pages of text. You are encouraged to include graphical and tabular summaries where appropriate (these do not count against the 6 pages of text) which can be included as an appendix.

You are free to write the summary as you wish, but one way to organize the written summary is in the following manner:

- Have your introduction describe the prediction task and the data, and follow this description with a summary of the final model(s) that you submitted for prediction along with the associated discrepancy measure (you can report the measure based on the 50% test data sample rather than the final measure).
- Describe the earlier and simple models you may have tried, and diagnostics or insights that led you to try other approaches. Briefly chronicle the main events that resulted in improvements in your predictive models. Here it might be interesting, for example, to keep track of the improvement in your discrepancy measure over time, and show the trajectory in a graph as a function of time.
- Report in some detail on the final models that resulted in your best predictions. Be clear about the model specifications or methods and justify your reasons for the various modeling decisions you made.
- Summarize the substantive conclusions of your final model. Which variables or combinations of variables were important predictors of voting? What groups of registered Colorado voters might you target if the goal was to identify unlikely voters?
- A critical evaluation of your overall approach can be insightful as well. What aspects of your modeling attempts did you expect would substantially improve predictive accuracy, but under-performed relative to what you anticipated? Where did you realize the greatest gains? In retrospect, what decisions could have made the process more efficient?

Please do not attempt to chronicle every step in the process that led to your final models, but instead focus on the key events or flashes of insight that resulted in the greatest improvement in predictability.

Project Grade

The project is worth 30% of your final course grade. If working on a team, all team members will receive the same project grade. From a grading perspective, the main criteria for a successful project include

- Evidence that you have learned material taught in the course. While you are encouraged to try statistical methods beyond those taught in the course for the prediction exercise, you should emphasize your experience using tools, methods, and concepts taught in the course, and incorporate them into the prediction exercise and your written summary.
- Evidence that you have put some time and thought into the project. Avoid rushing through the project as this will produce sloppy results. Because you are allowed at most five uploads per day, it would be a mistake to cram the work for this project into the day or two before the contest is over as you will have too few opportunities to get feedback on the success of your predictive modeling efforts.
- Clarity of your written summary and correctness of the content. When writing your summary, you should make sure your explanations are clear, and that you are using correct notation and terminology in describing your modeling and methods used. Your notation and terminology should be consistent with that developed in Stat 149 this semester.
- You will <u>not</u> be graded on the accuracy of your best predictions, nor your placement on the leaderboard in the Kaggle competition. On the other hand, if your best predictions do not outperform the simple benchmarks I post (such as using the sample fraction of those who vote in the *train.csv* file as the prediction for all 39,510 potential voters in the *test.csv* file), then this will raise questions about your level of effort in carrying out the project.