Statistics 149 — Spring 2018 — Course Project

Mark E. Glickman

Key information and milestones:

Decision to work alone or as a team: 5pm on Friday, March 30, 2018

Prediction contest: Ends 10:00pm on Sunday, April 29, 2018

Written report: Due 10:00pm on Wednesday, May 2, 2018

Prediction contest web site: `https://www.kaggle.com/c/political-leaning-prediction`

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://www.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 *www.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 four 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:

I will be making available a special URL that will permit you to enter the competition. You will need to set up an account on the *www.kaggle.com* web site. Please follow the instructions to sign up if you do not already have an account. 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 30. 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 voter self-reported support for the Democratic party candidate in 2016. These data were kindly provided by `bluelabs.com`. By clicking on the "Data" link on the contest page, you will be able to download two files. The first, *train.csv*, contains a randomly selected 10,440 observations (75%) from the data set of voting-eligible citizens. The data fields appear on the data page: `https://www.kaggle.com/c/political-leaning-prediction/data`. The last variable, `suppdem`, is the binary response, and the other variables are predictors of this binary response. Many of the variables (variables 15 through 45) were derived from proprietary models based on phone surveys and information from public voter files. It is worth mentioning that some of the variables contain missing data. You may want to investigate strategies for addressing the missing values.

The second file, *test.csv*, contains 3480 observations (25% of the original data set) with the same variables as above but with the variable `suppdem` withheld, and with the variable `Id` added.

Your job is to apply the model you developed on *train.csv* to obtain probability predictions on the withheld `suppdem` 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 3480 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 (called `sample-submission.csv`). You can submit your prediction file from the "Submit Predictions" link. 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. You will get to choose which two of your submissions you want scored for final placement in the competition - by default, Kaggle will choose the two with the best score on the 50% subset. 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 Kaggle 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. 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 3480 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, $i = 1, \cdots, n$, $(n = 3480)$ in the test data set, the total discrepancy is computed as

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

where y_i is the true (but withheld) indicator of whether potential voter i in the test data set self-reported supporting the Democratic candidate. 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 $n = 1740$ 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, 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. The six pages can be single-spaced if you like, but please use a font size no smaller than 11pt. 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. Attaching code is not necessary, but you may find it helpful to insert an occasional code snippet if it helps illustrate particular analyses you performed.

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.

- 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 modeling. Which variables or combinations of variables were important predictors of voting? You should plan to devote roughly 1/3 to 1/2 of your report to summarizing the final model and providing useful conclusions resulting from your analyses.
- 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. It is more important to focus on what the final models are telling you about the predictability of the response.

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, not with another course that used different notation.
- You will not 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, then this will raise questions about your level of effort in carrying out the project.