Final Project

(Due Saturday, December 16, 2017 at 11:59pm)

ISOM 674, Fall 2017

The final project is team based. That is, one submission should be made per team.

This project involves predicting clicks for on-line advertisements. The training data consists of data for 9 days from October 21, 2014 to October 29, 2014. The training data is in the file "ProjectTrainingData.csv." There are in excess of 30 million records in this file. The variables in this file are as follows:

- id = the identifier of the ad (this may, or may not be unique to each row).
- click = 1 means the ad was clicked on. click = 0 means the ad was not clicked on.
- hour = the date and hour when the ad was displayed. Format is YYMMDDHH.
- C1 = an anonymized categorical variable.
- banner_pos = the position in the banner.
- site_id = an identifier for the web site.
- site_domain = an identifier for the site domain
- site_category = a code for the site's category.
- app_id = an identifier for the application showing the ad.
- app_domain = an identifier for the app's domain.
- app_category = a code for the category of the app.
- device_id = an identifier for the device used.
- device_ip = a code for the ip of the device.
- device_model = the model of the device.
- device_type = the type of the device.
- device_conn_type = the type of the device's connection
- C14 – C21 = anonymized categorical variables

Thus, there are 24 columns in the dataset. The variable "click" is the Y-variable in the dataset. You will be attempting to predict the probability of a click.

The file "ProjectTestData.csv" contains the data that will be used to evaluate the performance of the approach you develop. It contains about 13 million records that are not a part of the "ProjectTrainingData.csv" data set. The columns in ProjectTrainingData.csv include all of the variables listed above except for "click." The actual clicks for these records are kept secret and will be used to evaluate the performance of your approach.

To submit your predictions, you will submit the file ProjectSubmission-TeamX.csv with the probabilities in the column labeled "P(click)" (which are currently al 0.5) changed to the probabilities predicted by your approach based on the explanatory variables in the file ProjectTestData.csv. Note that the order of the rows in the files ProjectTestData.csv and ProjectSubmission-TeamX.csv are the same and must be kept the same (otherwise you will be making predictions for the wrong rows). Also, you cannot assume that the "id" variable uniquely

identifies the rows. When you submit the file "ProjectSubmission-TeamX.csv" you will change the "X" to your team number (e.g., ProjectSubmission-Team10.csv).

The performance criterion used to evaluate the performance of your prediction is log-loss:

$$\text{Log Loss } = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right)$$

Note that the log loss is just the negative of the average of the terms in the log-likelihood. So, the smaller the log loss, the better. Here the y_i's correspond to the click variable (kept secret for the test data) and the p_i's are the predicted probabilities provided by your approach.

The deliverables for the project are: (1) The ProjectSubmission-TeamX.csv file containing your predicted probabilities (presumably based on the corresponding explanatory variables in the file ProjectTestData.csv); (2) A pdf document not to exceed 12 printed pages describing your approach (in a file named ProjectDescription-TeamX.pdf); and (3) all code files you used the project.

Please submit a zip file containing the deliverables called Bus674-Project-TeamX.zip which will unpack into a folder called TeamX. The TeamX folder should contain the files ProjectSubmission-TeamX.csv and ProjectDescription-TeamX.pdf and a folder "code" containing the code files. Of course, in every occurrence of TeamX in the above, replace the X by your team number.

Your code files should be written in R or Python and be appropriately documented. I do not expect to try to run your code files, but will briefly examine them to get a sense of how competently written they are.

Your pdf document submission ProjectDescription-TeamX.pdf should begin with your team number and the names of all team members. At the end should be an appendix listing out all of the code files and briefly indicating what they do. The document should be easily readable (i.e., 12pt) but otherwise the format is up to you.

The grading of the project will be based 60% on the performance of your predictions and 40% based on the description of your approach.

Honor Code: This is a team-based project so it can be discussed with your team members, but not with others. I did not generate this data set, so it might be possible to "find it" on the internet. Specifically searching for this data set or reading about analysis of it violates the honor code. The honor code includes a duty to report violations you observe. Finally, I will tell you that I have changed the data set from the original in ways that will make reliance on the analysis of others a risky business.