Course Project: Part 1

Due: **February 21, 2018, 5:00 PM**, via Blackboard

For the first part of the course project there are two objectives:

a) Choose a dataset.

b) Investigate and explore the dataset.

1 Choosing a dataset

You are strongly encouraged to choose a dataset in an area you find interesting and are motivated to explore. You have to like your data – you're going to spend a lot of hours staring at it, so you should find it fun and interesting to work with the dataset you've chosen!

Most zip files include dictionaries with interpretations of field names in addition to the raw data. Otherwise, the field names should be fairly self-explanatory. If anything is unclear, please ask.

2 Investigating and exploring the dataset

Once you've chosen your dataset, work out the following steps:

(a) Describe the dataset you have selected. Explain how the data was collected, and explain the meaning of the columns. Do you have any concerns about the data collection process, or about the completeness and accuracy of the data itself?

Note: This is also a good time to go through some basic *data cleaning*: if there are columns that are obviously extraneous to data analysis (e.g., IDs or metadata that have no bearing on your analysis), you can remove those now to make your life easier.

(b) Are any values in your dataset `NULL` or `NA`? Think of what you will do with rows with such entries: do you plan to delete them, or still work with the remaining columns for such rows? (You don't need to report anything to us for this part.)

(c) *Randomly* choose a test set (representing 20% of your rows), and keep it for later. You will not touch this test set again until the end of the course! Fix this set from the beginning, and use the remaining 80% for exploration, model selection, and validation. (You don't need to report anything to us for this part.)

Note: It may be harder to do this properly with some type of datasets, like time series; if you have selected time series data, let us know and we can help you find a testing strategy. In general, for such data, you want to train on earlier data and test on later data.

(d) Compute the mean and variance for each of the columns (you don't need to report this to us). Are there any columns that appear to be random noise?

(e) Suggest at least one possibility for a continuous outcome variable (a.k.a. response variable) that may be of interest to measure the effect of a potential intervention. Explain your choice and the variable's potential meaning. Compute the mean and variance of this variable.

(f) Suggest at least one possibility for a binary outcome variable. Compute the mean of this variable (i.e., the fraction of rows for which this variable is 1.) Explain your choice.

Note: You can always create a binary outcome variable by starting with a continuous variable Y, and then defining $Z = 1$ if Y exceeds a fixed threshold, and $Z = 0$ otherwise.

(g) Find the five covariates that are the most strongly positively correlated, as well as most strongly negatively correlated, with your choice of continuous outcome variable. (You should do the same for your choice of binary outcome variable, but you don't need to report the results.)

Are there variables you think should affect your outcome variable but actually have weak correlation with your outcome variable?

(h) Now find mutual correlations among the ten variables you identified in the last part. Create a scatterplot for every pair of covariates you believe correlates well to the outcome variable.

Are correlations associative in your data? That is, if A is correlated strongly with B, and B with C, is A also correlated strongly with C in your data?

(i) Are there variables you would like to *add* to your dataset as you embark on your analysis? For example, are there interactions or higher order terms that might be relevant? (You don't need to report your answer to this part.)

Final note: These steps are just the tip of the iceberg! Ideally, you will look at your data many different ways; for example, its useful to look at means and variances of columns, grouped based on the level of a categorical variable.

Try to play with and understand your data as much as you can *before* you start building models!