Analyze This!

Crowd-sourced Data Science

Bootcamp #1, Data-Preparation *(Saturday, 2/11/17, 8:00 AM to 12 Noon, Carlson School 1-135)*

0. Prerequisites. a) Member of AT!, b) Fully executed data-sharing agreement, c) Laptop with XL or Python.
1. Cost. $40 cash paid at the door.
2. The Challenge. Develop an algorithm to predict the 2nd gift potential of first-time donors to TCH4H
   a. Model 1. Predict Likelihood to give a 2nd gift.
   b. Model 2. Predict 2nd gift amount in dollars.
   c. $core = Model 1 * Model 2, dollars
   d. How will the $cores and BI get used by TCH4H?
3. How does the "question" influence the data preparation?
   a. Time frame …
      i. Time between 1st and 2nd gift (i.e., where did 3.25 years come from?)
      ii. Earliest date of RE "good" information
   b. Chicken & Egg. Limited to information available at first gift entry.
4. The data Dictionary. If none, how to create one.
5. Data stitching. Combining data from multiple sources.
6. Tidy the data …
   a. Checking for duplicates,
   b. Counting rows and columns,
   c. Missing cells (blank, NA, Unknown, Zero, etc.). Pros & Cons of imputation.
7. Data exploration …
   a. Continuous vs. Discrete
   b. Operational definitions
   c. Measurement variation
   d. Summarize & Graph via Pivot Tables, Frequency distributions and Trend charts.
8. Data conversion to Features …
   a. Continuous.
   b. Discrete. Does the rule of 5 apply? Don't check your brains at the door!
   c. Y is Binary. Special considerations and the rule of 5 (event & non-event both >= 5).
   d. Transformations. Non-linear, normalizing, scaling.
   e. Caution! Don't average or aggregate. But if you must …
9. 3rd Party additions ("wrangling") …
   a. Google API's
   b. Scraping web pages
10. Up Next. Modeling via Linear Regression and Gradient Boosted Decision Trees (3/11/17, 8A-5P, $40/$40).