MSIA 401 Project (Fall 2016)
Report Due: Wednesday, December 7
Professor Ajit Tamhane

- **Business Problem**: A non-profit organization that uses direct mail to solicit contributions from past donors wants to develop a predictive model to predict who will donate and how much in future.

- **Data**: A total of 99,200 data records are available, one per past donor. The data file `donation data.csv` contains the raw data for all records. There are 25 variables, which are listed in Table 1 and the summary statistics for them are given in Table 2. The base time period for all variables is from 10/2001 through 6/2010 except the outcome variable `TARGDOL`, which is the amount donated in the test time period 10/2010-12/2010 in response to a solicitation sent to all persons in the data file in 10/2010. If `TARGDOL` = 0 then the person did not donate. The records are listed in the decreasing order of `TARGDOL` with the first 27,208 records with `TARGDOL` > 0 and the remaining 71,992 with `TARGDOL` = 0. Thus the response rate is 27%. All other variables (except ID, of course) are possible predictors.

  Some comments about the data.

  1. There are many missing values. Missing values should not be automatically replaced by an average of non-missing values for every variable. For example, CNDOL1 is the amount of the latest contribution and CNDOL2 is the previous donation, but people who have only donated once have CNDOL2 missing. In this case CNDOL2 should be set equal to 0. For `SEX` variable, U means sex is not reported and this cannot be replaced; B means a married couple, not sure what C means.
  2. Contributions codes (CNCOD1, CNCOD2, CNCOD3) and solicitation codes (SLCOD1, SLCOD2, SLCOD3) are four digit codes, which are grouped into five types (A, B, C, D and M) in the file `dmef1code.csv`. Any other type implies missing values. Do not impute missing values since they mean that a there was no donation. I don't know what "type" exactly means but presumably it represents some type of incentive (e.g., free address labels). You will need to merge the `donation data.csv` file with the `dmef1code.csv` to map the four digit codes to the five types and use the "type" as a categorical variable. The `STATE` variable has too many categories to be a useful predictor, unless you group the states into 7 or 8 regions of the country.
  3. Data are mostly missing on CNCOD2, CNCOD3 and SLCOD2, SLCOD3 since these people donated only once. It is not possible to impute their missing values, you may choose to omit these variables. CNCOD and SLCOD variables essentially give the same information. Consider for example, the first observation in the data file (ID=11296). This customer donated most recently during month 181 (cndat1), which is 5 months ago (cnmon1), so the current month must be 186. The previous donation was during month 169 (cndat2) which is $186 - 169 = 17$ months ago (cnmon2). The third most recent donation was during month 157 (cndat3), which is $186 - 157 = 29$ months ago (cnmon3). If you look at the next row (ID=27046), you see that the first donation (cnmonF) came 44 months ago, which matches cnmon2, since this person has only donated twice and cnmon3 is missing. The most recent donation was in month 149 (cndat1), which was $186 - 149 = 37$ months ago (cnmon1).

- **Suggested Methodology**:

  1. Based on preliminary analyses, transform the data and include any interactions as appropriate.
  2. Divide the data into training and test sets in the ratio of 2:1 by putting every 3rd observation in the test set and all remaining observations in the training set. This is to insure that all student

groups use the same training and test sets. All model fitting should be done on the training set and model validation on the test set.

3. First develop a binary logistic regression model for `TARGDOL` > 0. This model will be used to estimate the probabilities of donating for the test set.
4. Next develop a multiple regression model using data with `TARGDOL` > 0 only.
5. For each observation (including `TARGDOL` $= 0$) calculate E(`TARGDOL`) by multiplying the predicted `TARGDOL` from the multiple regression model by P(`TARGDOL` > 0) from the logistic regression model by using the formula $E(y) = E(y|y > 0)P(y > 0)$.
6. Compute root mean square prediction error for `TARGDOL` for the test set.
7. Select 1000 donors from the test set who have the highest E(`TARGDOL`). These may be the donors that will be special marketing targets. Then find their total actual donations. This is the payoff and should be as high as possible.

- **Written Report**

  The text part of the report should be no more than 20 pages (double spaced, 12 point font). Put the outputs, plots etc. in an appendix. The report should roughly follow the outline below.

  1. Cover page (Title, names of group members)
  2. Executive Summary: Give a non-technical summary of your findings mentioning the key predictors of responders vs. non-responders and of the amount of contributions. This summary is for upper management and should not include any equations and as few statistics as possible. So don't mention things like R2 here. (About 1/2 page)
  3. Introduction: Describe your overall approach and any a priori hypotheses. Give a brief outline of the other sections of the report. (About 2 pages)
  4. Model Fitting: This is the core of the report. Divide this into two parts: (i) logistic regression model, (ii) multiple regression model. Explain the steps used in model fitting including exploratory analysis of data to assess the nature of relationships, detection of outliers and influential observations, linearizing and normalizing transformations etc.; different models fitted and methods used to fit them (e.g., stepwise regression); model diagnostics. The final model including residual analyses and other diagnostics resulting to data transformations. (About 12 pages)
  5. Model Validation: Explain how you validated the model against the test data set. Report the results about how well the model predicted the test set sales values and how well your top 1,000 predicted customers performed in terms of contributions. (About 4 pages)
  6. Conclusions: Draw conclusions about significant predictors, any key missing predictors which would have improved the model, etc. (About 2 pages)
  7. References
  8. Appendix (Printouts, Graphs)

Table 1: Description of Variables

```
---------------------------------------------------------------
 #    Variable    Type    Len    Label
 --------------------------------------------------------------
 1    CNDOL1      Num       6    Latest Contribution
 2    CNTRLIF     Num       6    Dollars Contribution Lifetime
 3    CONLARG     Num       6    Largest Contribution
 4    CONTRFST    Num       6    First Contribution
 5    CNCOD1      Num       4    Latest Contribution Code
 6    CNCOD2      Num       4    2nd Latest Contribution Code
 7    CNCOD3      Num       4    3rd Latest Contribution Code
 8    CNDAT1      Num       4    Latest Contribution Date
 9    CNDAT2      Num       4    2nd Latest Contribution Date
10    CNDAT3      Num       4    3rd Latest Contribution Date
11    CNDOL2      Num       4    2nd Latest Contribution
12    CNDOL3      Num       4    3rd Latest Contribution
13    CNTMLIF     Num       4    Times Contributed Lifetime
14    SLCOD1      Num       4    Latest Solicitation Code
15    SLCOD2      Num       4    2nd Latest Solicitation Code
16    SLCOD3      Num       4    3rd Latest Solicitation Code
17    TARGDOL     Num       4    Dollars of Fall 1995 Donations
18    STATCODE    Char      2    State
19    SEX         Char      1    Gender
20    CNMON1      Num       3    Months since latest contrib
21    CNMON2      Num       3    Months since latest 2nd contrib
22    CNMON3      Num       3    Months since latest 3rd contrib
23    CNMONF      Num       8    Months since first contrib
24    CNMONL      Num       8    Months since largest contrib
25    ID          Num       4
----------------------------------------------------------------
```

Table 2: Summary Statistics

Variable	Label	N	# Missing	Mean	Minimum	Maximum
CNDOL1	Latest Contribution	99200	0	9.327	2.00	1000.00
CNTRLIF	Dollars Contribution Lifetime	99200	0	33.199	1.00	4440.00
CONLARG	Largest Contribution	99200	0	10.535	2.00	1000.00
CONTRFST	First Contribution	99200	0	8.000	0	500.00
CNCOD1	Latest Contribution Code	99200	0	6345.86	18.00	718.00
CNCOD2	2nd Latest Contribution Code	39775	59425	5347.76	83.00	692.00
CNCOD3	3rd Latest Contribution Code	31798	67402	4674.63	7.00	7242.00
CNDAT1	Latest Contribution Date	99200	0	177.018	60.00	186.00
CNDAT2	2nd Latest Contribution Date	39775	59425	164.773	48.00	186.00
CNDAT3	3rd Latest Contribution Date	31798	67402	154.693	36.00	185.00
CNDOL2	2nd Latest Contribution	39775	59425	9.296	1.00	750.00
CNDOL3	3rd Latest Contribution	31798	67402	8.958	1.00	600.00
CNTMLIF	Times Contributed Lifetime	99200	0	3.983	1.00	55.00
SLCOD1	Latest Solicitation Code	99200	0	7313.83	998.00	7704.00
SLCOD2	2nd Latest Solicitation Code	74799	24401	6701.30	998.00	7683.00
SLCOD3	3rd Latest Solicitation Code	57182	42018	6326.85	998.00	7312.00
TARGDOL	Dollars of Fall 1995 Donations	99200	0	2.325	0	1500.00
CNMON1	Months since latest contrib	99200	0	8.982	0	126.00
CNMON2	Months since latest 2nd contrib	39775	59425	21.227	0	138.00
CNMON3	Months since latest 3rd contrib	31798	67402	31.307	1.00	150.00
CNMONF	Months since first contrib	99200	0	34.947	0	1146.00
CNMONL	Months since largest contrib	99200	0	16.656	0	162.00
ID		99200	0	49600.50	1.00	99200.00