MSIA 401 Project (Fall 2017)
Report Due: Friday, December 1

- **Business Situation**: The file `catalog sales data.csv` posted on Canvas under Data Sets page comes from a retail company that sells upscale clothing on its website and via catalogs, which help drive customers to the website. All customers were sent a catalog mailing on Sep 1, 2012. On Dec 1, 2012 it was recorded whether or not they responded by making a purchase. There is one row for each customer. The `targdol` is the response variable, which is the purchase amount in response to receiving the catalog (`targdol` = 0 indicates that the customer did not respond). The remainder of variables are potential predictor variables which give information about the customer as of the time of the mailing. LTD means "life-to-date," i.e., since the customer purchased for the first time.

- **Data**: There are a total 101,532 customers, who are randomly split into 50418 in the training set and the remaining 51,114 in the test set (`train` =1 training set, `train` =0 test set). The definitions of the variables are as follows.

  - `targdol`: dollar purchase resulting from catalog mailing
  - `datead6`: date added to file
  - `datelp6`: date of last purchase
  - `lpuryear`: latest purchase year
  - `slstyr`: sales ($) this year
  - `slslyr`: sales ($) last year
  - `sls2ago`: sales ($) 2 years ago
  - `sls3ago`: sales ($) 3 years ago
  - `slshist`: LTD dollars
  - `ordtyr`: orders this year
  - `ordlyr`: orders last year
  - `ord2ago`: orders 2 years ago
  - `ord3ago`: orders 3 years ago
  - `ordhist`: LTD orders
  - `falord`: LTD fall orders
  - `sprord`: LTD spring orders
  - `train`: training/test set indicator (1 = training, 0 = test)

- **Goal**: Build a predictive model for `targdol` based on the training set and then test it on the test set.

- **Strategy for Building the Prediction Model**: Only 10% of the customers were responders (had `targdol` > 0). So straightforward multiple regression will not work. You need to adopt a two-step model fitting approach.

  1. Based on preliminary analyses, transform the data and include any interactions as appropriate.

2. First develop a binary logistic regression model for `targdol` > 0. Use this model to estimate the probabilities of being responders for the test set.
3. Next develop a multiple regression model using data with `targdol` > 0 only.
4. 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 $E(y) = E(y|y > 0)P(y > 0)$.

- **Criteria for Evaluating the Fitted Models**: The final fitted regression model should meet the usual criteria such as significant coefficients, satisfactory residual plots, good fit as measured by R^2 or R^2_{adj}, parsimony and interpretability of the model etc.

  Two numerical criteria will be used to evaluate the fitted models on the test set.

  Statistical Criterion : Mean square prediction error (MSPE) $= \left[\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\right] / [n-(p+1)]$; $p+1$ is the number of β coefficients in the multiple regression model derived from the training set.

  Financial Criterion : Select the top 1000 customers (prospects) from the test set who have the highest E(`targdol`). Then find their total <u>actual</u> purchases. This is the payoff and should be as high as possible.

- **Hints**:

  1. These data are dirty and you will have much cleaning up to do. Some errors in the data are as follows. If you run a histogram or frequency distribution of the `datelp6` variable among only those with `targdol` > 0 you will see that, for the most part, `datelp6` equals one of two distinct dates in the calendar year. It is as if the person who prepared the data did some strange rounding within six-month bins. There are also other inconsistencies in the data, e.g., `falord` + `sprord` is not equal to `ordhist` in about 9% of the cases Similarly, the year of latest purchase obtained from `lpuryear` variable and from `datelp6` variable do not always agree. Some of these errors result because when two variables measure the same thing, both are not updated.
  2. It is known in data base marketing that generally the best predictors for deciding whether a customer will respond to a catalog are (1) recency of last purchase and (2) consistency of past purchases. Recency can be readily deduced from the date of last purchase. Consistency can be coded as an interaction of the last 1, 2 or 3 years of sales or orders. You would need to create such interaction variables.
  3. The significant predictors for the classification model will be generally different from those for the multiple regression model.