MSIA 401 Project (Fall 2018)
Report Due: Thursday, December 6, 5 PM

- **Business Situation**: A German online book seller has provided data on a sample of 33,713 customers on their purchases of books prior to 01AUG2014 when a promotional offer was made and their purchase amounts (`targamt`) in euros over the next 3 months in response to the offer. The total sample of 33,713 customers is divided into a training sample of 8311 customers and a test sample 25,402 customers. The goal of the project is to build a predictive model for `targamt` based on the predictor variables from the past purchase history and use the model to predict `targamt` for the test sample customers. The response rates (proportion of customers with `targamt` > 0, i.e., who bought a book) are $327/8311 = 3.93\%$ in the training sample and also $999/25402 = 3.93\%$ in the test sample. Note that `targamt` is log-transformed as `logtargamt` $= \ln($`targamt`$+1)$, so that if `targamt` $= 0$ then `logtargamt` is also 0.

- **Data**: There are a four data files which can be matched by common customer id's.

  1. `book.csv` file: This is the most extensive data file with the following data on all 33,713 customers.

     `id`: unique customer id

     `logtargamt`: blank for the test sample

     `recency`: no. of days since the last order

     `frequency`: number of orders

     `amount`: total past purchase amount in euros (not sure why all purchase amounts and prices are reported to many decimal places)

     `tof`: time on file

     `Fxx`: frequency of orders of books of category xx

     `Mxx`: amount of purchase of books of category xx

     The following are the categories : 1=fiction, 3=classics, 5=cartoons, 6=legends, 7=philosophy, 8=religion, 9=psychology, 10=linguistics, 12=art, 14=music, 17=art reprints, 19=history, 20=contemporary history, 21=economy, 22=politics, 23=science, 26=computer science, 27=traffic, railroads, 30=maps, 31=travel guides, 35=health, 36=cooking, 37=learning, 38=games and riddles, 39=sports, 40=hobbies, 41=nature/animals/plants, 44=encyclopedias, 50=videos, DVDs, 99=non-books

     The aggregate `frequency` and `amount` variables are the totals of `Fxx` and `Mxx` variables, respectively. I believe only the aggregate variables are important, so `Fxx` and `Mxx` variables may be ignored.

  2. `ordersall.csv` file: This file contains data on all 627,955 orders, which translates to an average of 18.45 orders per customer. The data fields are as follows.

     `id`: unique customer id

     `orddate`: order date

     `ordnum`: order number

     `category`: category of the book

     `qty`: quantity ordered

     `price`: price

  3. `booktrain.csv`: This file has only two variables: `id` and `logtargamt` for 8311 customers in the training set.

4. `booktest.csv`: This file also has only two variables: `id` and `logtargamt` for 25,402 customers in the test set.

- **Strategy for Building the Prediction Model**: Only 3.93% of the customers are responders. So straightforward multiple regression will not work. You need to adopt a two-step model fitting approach.

  1. Based on preliminary analyses, transform any variables (note `targamt` is already log-transformed). You can create additional feature variables, e.g., orders and purchase amounts per unit time on file. Instead of total past purchase amount, you might consider discounting the older purchases more than the more recent ones.
  2. First develop a binary logistic regression model for responders. Use this model to estimate the probabilities of being responders for the test set.
  3. Next develop a multiple regression model using the training set data for the responders only, i.e., those customers with `logtargamt` > 0.
  4. For each observation (including those with `logtargamt` $= 0$) in the test set calculate E(`logtargamt`) by multiplying the predicted `logtargamt` from the multiple regression model by P(`logtargamt` > 0) from the logistic regression model by using the formula $E(y) = E(y|y > 0)P(y > 0)$. This gives the predicted `logtargamt` for the test set customers. Calculate the `targamt` amounts from these by exponentiating and by subtracting 1 (reverse of the log-transformation).

- **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.

  Use the following two numerical criteria to evaluate the fitted models on the test set.

  Statistical Criterion : The sum of squared errors of prediction (SSEP) obtained by summing the squares of the differences between the actual `targamt` and the predicted `targamt` values for the test sample.

  Financial Criterion : Select the top 500 customers (prospects) from the test set who have the highest E(`targamt`). Then find their total actual purchases. This is the payoff and should be as high as possible. What percentage is this of the total purchases of the actual top 500 customers in the test sample? This percentage should be as high as possible.

  Instead of fixing the number of top prospects at 500, you might try to find the optimum number of top prospects by maximizing the short term profit if the profit margin is 25% of `targamt` and the cost of mailing the promotional material to each prospect is 1 euro. Thus you will have to find x to maximize 0.25*(sales revenue from the top x prospects)-1*x.