**MSIA 401 Project**
**Report Due: Friday, December 4**
**Professor Ajit Tamhane**

- **Business Situation**: A speciality, multichannel retailer sells to customers on its website and via catalogs, which drive customers to the website. All customers were sent at least one catalog during the the fall of 2008. All purchases made during the fall of 2008 are recorded in the `TARGAMNT` variable. All other variables capture some aspect of their behavior prior to the fall of 2008 and should be used as predictor variables.
- **Goal**: Build a predictive model for `TARGAMNT`
- **Data**: You are given two data sets, `training.csv` and `test.csv`. You should build your model using `training.csv`. After you have built your final model, apply it to `test.csv`, which has the same variables as `training.csv`. Here are descriptives for the training set:

| Variable | Label | N | Minimum | Maximum |
|---|---|---|---|---|
| RECMON | Months since last order | 52844 | 0 | 61.000 |
| ORDCLS1 | 5 Year Product Class 1 Orders | 52844 | 0 | 5.000 |
| ORDCLS2 | 5 Year Product Class 2 Orders | 52844 | 0 | 10.000 |
| ORDCLS3 | 5 Year Product Class 3 Orders | 52844 | 0 | 6.000 |
| ORDCLS4 | 5 Year Product Class 4 Orders | 52844 | 0 | 3.000 |
| ORDCLS5 | 5 Year Product Class 5 Orders | 52844 | 0 | 6.000 |
| ORDCLS6 | 5 Year Product Class 6 Orders | 52844 | 0 | 9.000 |
| ORDCLS7 | 5 Year Product Class 7 Orders | 52844 | 0 | 8.000 |
| SALCLS1 | 5 Year Sales Product Class 1 | 52844 | -91.820 | 678.040 |
| SALCLS2 | 5 Year Sales Product Class 2 | 52844 | -127.350 | 2232.31 |
| SALCLS3 | 5 Year Sales Product Class 3 | 52844 | -80.870 | 3306.09 |
| SALCLS4 | 5 Year Sales Product Class 4 | 52844 | -65.150 | 1401.19 |
| SALCLS5 | 5 Year Sales Product Class 5 | 52844 | 0 | 1274.70 |
| SALCLS6 | 5 Year Sales Product Class 6 | 52844 | -114.970 | 1778.65 |
| SALCLS7 | 5 Year Sales Product Class 7 | 52844 | -87.250 | 5536.60 |
| ORD185 | Order Yr 1, Prom 85 (Y/N) | 52844 | 0 | 1.000 |
| ORD285 | Order Yr 2, Prom 85 (Y/N) | 52844 | 0 | 1.000 |
| ORD385 | Order Yr 3, Prom 85 (Y/N) | 52844 | 0 | 1.000 |
| ORD485 | Order Yr 4, Prom 85 (Y/N) | 52844 | 0 | 1.000 |
| TOF | Time on File | 52844 | 1.367 | 301.970 |
| TOTORD | Lifetime Orders | 52844 | 1.000 | 123.000 |
| TOTSALE | Lifetime Sales | 52844 | -127.350 | 7213.52 |
| TARGAMNT | Prom 85 Sales in Targ Wndw | 52844 | -18.550 | 1053.57 |

- **Hints**:
  1. Most of the predictor variables are counts or amounts and are right skewed with outliers. You should probably transform them as necessary.
  2. About 95% of the values of targamnt are 0 (the customers did not make a purchase). First build a binary logistic regression model to predict whether a customer made a purchase or not.
  3. Then build a multiple regression model for targamnt (or `logtargamnt`) for those cases where **targamnt**$> 0$. The predictor variables for the multiple regression model may be different from those for the binary logistic regression model.

4. You should consider creating new predictor variables from the given set. For example, `AOA` = average order amount = `TOTSALE/TOTORD` or `PR` = purchase rate = `TOTORD/TOF` may be good predictors of `targamnt` in the multiple regression model (when `targamnt`>0). (Note that TOF measures how long someone has been a customer.) There may be other "interaction" variables that are good predictors.
5. To find the predicted value of targamnt for each customer in the test set, multiply the predicted probability for the customer from the binary regression model with the predicted targamnt from the multiple regression model.
6. You do not have to follow this approach.

- **Evaluation**: I will evaluate your models on two criteria:

1. Compute average prediction error for targamnt for the test set (both squared error and average signed error)
2. Select 1000 customers who have the highest predicted sales. The team that makes the most money (adding up their actual sales of the top 1000 customers) wins this criterion.