**Carnegie Mellon University, Heinz College**

**Business Intelligence & Data Mining with SAS Suite**

**Fall 2012 (94832 Mini 1)**

**Assignment 3 - Predictive Modeling (Regression)**

**(Due Oct/9/2012)**

(Continue the Organics diagram from your HW2. A supermarket is offering a new line of organic products. The supermarket’s manager wants to determine which customers are likely to purchase these products in order to conduct future promotions.) **Now that you have helped him build a Decision Tree model for predictive analytics, he is quite happy with the outcome. He decides to promote you to be his chief business intelligence consultant (with an even higher salary) to conduct further analysis.**

The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of their loyalty program participants and collected data that includes whether or not these customers purchased any of the organic products. The **ORGANICS** data set contains over 22,000 observations and 18 variables. The variables in the data set are shown below with the appropriate roles and levels. Note that although two target variables are listed, this analysis concentrates on only the binary variable **ORGYN**.

****

**Questions:**

**(Note: It is recommended that you provide screen shots of your analysis along with the discussion.)**

*(Same as in HW2). Reuse the diagram Organics. Define the data set AAEM.ORGANICS (Metadata Repository--> Shared Data --> Libraries --> AAEMSPEL--> Organics) as a data source for the project.* *Set the model role for the target variable and examine the distribution of the variable.*

**a.** Explore the **ORGANICS** data source. In preparation for regression, is any missing values imputation needed? Why or why not? If yes, what should you do before generating the regression models? (Add an Impute node to the diagram and connect it to the Data Partition node...)

**b.** Change Default Input Method to Tree for both class and interval variables. Create missing value indicator variables. Replace missing values for **GENDER** with **U** for unknown.

**c.** Add a Regression node to the diagram and connect it to the Impute node. Choose the stepwise selection and average squared error as the selection criterion. Run the Regression node and view the results. Which variables are included in the final model? Which variables are important in this model?

**d.** In preparation for regression, are any transformations of the data warranted? Why or why not? (Add a Transform Variables node to the diagram and connect it to the Impute node...)

**e.** The variable **AFFL** appears to be skewed to the right. Use a square root transformation for **AFFL**. The variables **BILL** and **LTIME** also appear to be skewed to the right. Use SAS Enterprise Miner to transform these variables to maximize normality. Run the Transform Variables node. Explore the exported training data. Did the transformation of **AFFL** appear to result in a less skewed distribution? What transformation should be chosen for the variables **BILL** and **LTIME**?

**f.** Add another Regression node to the diagram and connect it to the Transform Variables node. Choose the stepwise selection method and average squared error selection criterion.Run this new Regression node and view the results. Which variables are included in the final model? Which variables are important in the model?

**g.** How do the validation average squared errors of the two regression models compare? Which one would you find a better model? Why?

**h.** Change the stepwise selection to forward selection and backward selection, respectively. Compare the three models. Which variables are included in each of the three models? Which variables are important in each of the three models? Which model provides the lowest validation average squared errors? Which model would you recommend to use? Why?

**i.** How do the regression models compare to the tree models that you built in HW2? Which one(s) would you prefer to use? Why?