Homework 7

Mark Van Wormer z1800639

1. **Creating, Exploring, and Preparing Data**

A supermarket offers a new line of organic products. The supermarket's management wants to determine which customers are likely to purchase these products.

The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of the loyalty program participants and collected data that includes whether these customers purchased any of the organic products.

The **ORGANICS** data set contains 13 variables and more than 22,000 observations. The variables in the data set are shown below with the appropriate roles and levels.

| **Name** | **Model Role** | **Measurement Level** | **Description** |
| --- | --- | --- | --- |
| **ID** | Key | Nominal | Customer loyalty identification number |
| **DemAffl** | Input | Interval | Affluence grade on a scale from 1 to 30 |
| **DemAge** | Input | Interval | Age, in years |
| **DemCluster** | Rejected | Nominal | Type of residential neighborhood |
| **DemClusterGroup** | Input | Nominal | Neighborhood group |
| **DemGender** | Input | Nominal | M = male, F = female, U = unknown |
| **DemReg** | Input | Nominal | Geographic region |
| **DemTVReg** | Input | Nominal | Television region |
| **PromClass** | Input | Nominal | Loyalty status: tin, silver, gold, or platinum |
| **PromSpend** | Input | Interval | Total amount spent |
| **PromTime** | Input | Interval | Time as loyalty card member |
| **TargetBuy** | Target | Binary | Organics purchased? 1 = Yes, 0 = No |
| **TargetAmt** | Rejected | Interval | Number of organic products purchased |

**🖉** Although two target variables are listed, these exercises concentrate on the binary variable **TargetBuy**. Be sure to select the role **Rejected** for **TargetAmt**. Also be sure to assign the role **Key** to **ID**.

* 1. Create a new diagram in your current project. Name the diagram HP Organics. Use the Chapter 4 library to define the **ORGANICS** data set as a data source for the project. If you did not set up the Chapter 4 library in the demonstrations, you would need to set it up now.
     1. Set the model roles for the variables as shown above.
     2. Why is the variable **DemCluster** rejected? Because it only told us the type of residential neighborhood which was not pertinent to our data set.
     3. Finish the **Organics** data source definition by defining the role of the data table as **Raw.**
  2. Add the **CHAPTER4.ORGANICS** data source to the HP Organics diagram. Add an HP Explore node to the diagram and connect the **Organics** data source to it. Run the HP Explore node and view the results.

Are there any variables with missing values? If so, which variables? Which has the highest percentage missing? DemAffl, DemAge, DemCluster, DemGender, DemReg, DemTVReg, and PromTIme. DemGender has the highest percent missing at 11.3036%

* 1. Add an HP Data Partition node to the diagram and connect the **Organics** data source to it. Perform a 50/50 partition for training and validation.

What is the approximate rate of organics purchasers (that is, 1s) in the training and validation data sets? For observation (1) there are 5505 observations 2759 from training, and 2746 from validation.

**Constructing Models**

The next portion of these exercises asks you to construct several predictive models. Some data cleansing is performed prior to training certain model types.

* 1. Add an **HP Forest** node to the diagram and connect the **HP Data Partition** node to the **HP Forest** node. In the Properties panel, change Maximum Number of Trees to **50** and use the remaining default settings to construct the forest. Run the HP Forest node and view the results.
     1. Which variable appeared in the most number of splitting rules? DemAffl at 317 splitting rules.
     2. What are the values of Average Squared Error (ASE) on training and validation for the forest? The values on the ASE are .13 and the validation is .14
  2. Add a second **HP Forest** node to the diagram and connect the **HP Data Partition** node to it. This forest will consist of a larger number of trees than the default forest, so rename the node **HP Forest More Trees**, but this second forest will also use a larger sample of observations for each tree. In the Properties panel, change Maximum Number of Trees to **100** and change Proportion of obs in each sample to **0.8**. Run the HP Forest More Trees node and view the results.
     1. Which variable appeared in the most number of splitting rules? DemAffl at 745 splitting rules
     2. What are the values of ASE on training and validation for this model? The values stay the same on the ASE are .13 and the validation is .14
  3. Based on validation ASE, which model appears to be better, the default forest or the forest based on more trees? The models appear to be similar with the change in the splitting rules I’d say the HP Forest More Trees is better.
  4. In preparation for regression modeling, extreme values should be investigated (and possibly transformed) and imputation must be performed. Add an **HP Transform** node to the diagram and connect the **HP Data Partition** node to it. Using the HP Transform node, graphically explore the four interval input variables.

Three of the four interval input variables show some degree of skewness. Which interval input variables are candidates for transformation? DemAffl, PromSpend, and PromTime

Apply a log transformation to the three input variables that are skewed. Run the HP Transform node, but do ***not*** view results.

* 1. Exploration of missing values occurred in an earlier step. Add an **HP Impute** node to the diagram and connect the **HP Transform** node to it. Set the node to impute a constant character value **U** for unknown class variable values and the overall mean for unknown interval variable values. Create imputation indicator variables for all imputed inputs and set the role of these indicators to **input**. Run the HP Impute node but do ***not*** view the results.
  2. Add an **HP Regression** node to the diagram and connect the **HP Impute** node to it. Select **Stepwise** for the Selection Method, **SBC** for the Selection Criteria, and **Significance Level** for the Stop Criteria. Run the HP Regression node and view the results.

What variables are included in the final model? TargetBuy, DemAffl, DemCluster, DemGender

What is the validation ASE? The Validation ASE is .14

* 1. In preparation for neural network modeling, in addition to the data preparation used for regression modeling, variable selection is required. Add an **HP Variable Selection** node to the diagram and connect the **HP Impute** node to it. Keep the default settings for the HP Variable Selection node. Run the HP Variable Selection node and view the results.
     1. What variables were rejected by **HPReduce** in the unsupervised part of the two-step variable selection method? TargetAMT, M\_LOG\_PromTime, M\_DemTVReg, M\_DemClusterGroup, and DemCluster.
     2. What variables are ultimately selected by the HP Variable Selection node to be passed on to the neural network? I would say the variables that are passed on would be
  2. Add an **HP Neural** node to the diagram and connect the **HP Variable Selection** node to it. Keep the default settings as they are in the Properties panel for the HP Neural node. Run the HP Neural node and view the results.

What is the validation ASE and how does it compare to the validation ASE from the three other models?

1. **Model Comparison and Scoring**

The final portion of these exercises consists of model comparison on the trained models and model deployment. A **score** data set is created before deployment.

* 1. On the Assess tab, drag a **Model Comparison** node onto the diagram. Connect the **HP Forest**, **HP Forest More Trees**, **HP Regression**, and **HP Neural** nodes to the **Model Comparison** node. Select **Average Squared Error** for the Selection Statistic and **Validation** for the Selection Table. Run the Model Comparison node and view the results.

Based on the **Fit Statistics** table, what model is selected? Both models return fairly similar results but based on judgement I would chose the Validate data.

* 1. Use the **ABA** library to define the **SCOREORGANICS** data source, which you use for scoring. Be sure to assign the role **Score** to the data source at the Data Source Attributes step. Score the **SCOREORGANICS** data source using the **Score** node and the model selected by the Model Comparison node.



