Homework 6

1. **Initial Data Exploration**

A supermarket is offering 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 their 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** | ID | 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**.

* 1. Create a new project named Homework and a diagram named Organics.
  2. Define the data set **Chapter3.ORGANICS** as a data source for the project.
* Set the roles for the analysis variables as shown above. Change the value of **Class Levels Count Threshold** to **2**.
* Examine the distribution of the target variable. What is the proportion of individuals who purchased organic products?

**Answer: .247 = .25 people bought organic products**

* + 1. The variable **DemClusterGroup** contains collapsed levels of the variable **DemCluster**. Presume that, based on previous experience; you believe that **DemClusterGroup** is sufficient for this type of modeling effort. Set the model role for **DemCluster** to **Rejected**.
    2. As noted above, only **TargetBuy** will be used for this analysis, and should have a role of **Target**. Can **TargetAmt** be used as an input for a model used to predict **TargetBuy?** Why or why not?

**Answer: TARGETAMT could be used but it would give much more information rather than indicating whether or not a customer purchased organics it indicates how much they ordered and its more cluttered.**

* + 1. Finish the **Organics** data source definition.
  1. Add the **Chapter3.ORGANICS** data source to the Organics diagram workspace.
  2. Add a **Data Partition** node to the diagram and connect it to the **Data Source** node. Assign 50% of the data for training and 50% for validation.
  3. Add a **Decision Tree** node to the workspace and connect it to the **Data Partition** node.
  4. Create a decision tree model interactively, automatically, or autonomously using average squared error as the model assessment statistic.
     1. Create the tree.
     2. How many leaves are in the optimal tree?

**Answer: 29 Leaves**

* + 1. Which variable was used for the first split? What were the competing splits for this first split?

**Answer: Affluence Grade**

1. **Predictive Modeling Using Regression**
   1. Return to the Organics diagram in the Exercises project. Use the **StatExplore** tool on the **ORGANICS** data source.
   2. In preparation for regression, is any missing values imputation needed? If yes, should you do this imputation before generating the decision tree models? Why or why not?

You should do this before generating a decision tree as it will affect the optimal tree and if there are any missing variables the optimal tree will be off

* 1. Add an Impute node to the diagram and connect it to the Data Partition node. Set the node to impute “U” for unknown class variable values, the overall mean for unknown interval variable values, and create imputation indicators for all imputed inputs.
  2. Add a **Regression** node to the diagram and connect it to the **Impute** node.
  3. Choose the stepwise selection and validation error as the selection criterion.
  4. Run the **Regression** node and view the results. Which variables are included in the final model? Which variables are important in this model? What is the validation ASE?

TargetBuy, P.TargetBuy, R.TargetBuy

TargetBuy is included in the final model as that is the amount indicator of how much organics were purchased.

Validation ASE = 0.32

* 1. In preparation for regression, are any transformations of the data warranted? Why or why not? I’d say the only transformation would be to change the variable from TargetAmt to TargetBuy as we would only like to indicate whether or not someone bought organic products.
  2. Disconnect the **Impute** node from the **Data Partition** node. Add a **Transform Variables** node to the diagram and connect it to the **Data Partition** node. Connect the **Transform Variables** node to the **Impute** node.
  3. Apply a log transformation to the **DemAffl**, **PromSpend**, and **PromTime** inputs.
  4. Run the **Transform Variables** node. Explore the exported training data. Did the transformations result in less skewed distributions? I would say there were less skewed distribution in the training data.
  5. Rerun the **Regression** node. Do the selected variables change? How about the validation ASE?

The variables do change and the validation ASE changed to 0.19

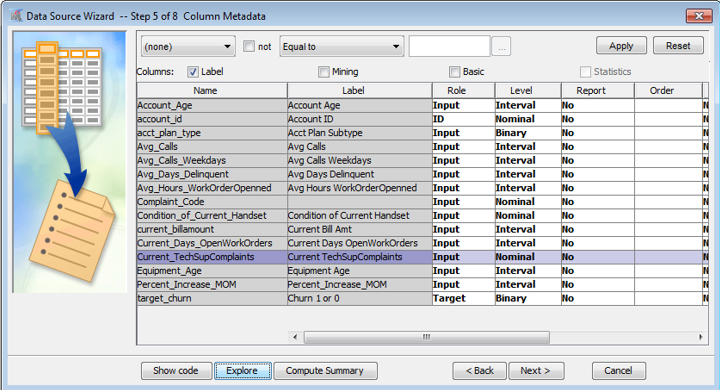
* 1. Create a full second-degree polynomial model. How does the validation average squared error for the polynomial model compare to the original model?

It compares to the other model as it changes to become smaller than the original model

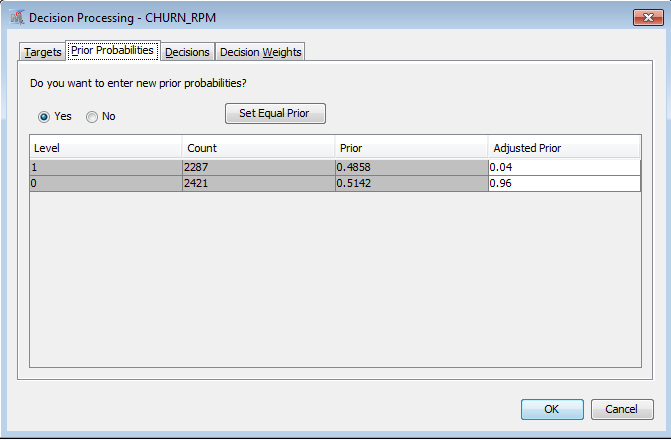
1. **Predictive Modeling Using Neural Networks**
   1. In preparation for a neural network model, is imputation of missing values needed? Why or why not?
   2. In preparation for a neural network model, is data transformation generally needed? Why or why not?
   3. Add a Neural Network tool to the Organics diagram. Connect the **Impute** node to the **Neural Network** node.
   4. Set the model selection criterion to average squared error.
   5. Run the **Neural Network** node and examine the validation average squared error. How does it compare to other models?
2. **Scoring Organics Data**
   1. Create a Score data source for the **ScoreOrganics** data.
   2. Score the **ScoreOrganics** data using the model selected with the Model Comparison node.
3. **Setting Up the Initial Project, Diagram, Library, and Data Source**

Churn refers to the tendency of a subscriber to switch providers; it is a common problem faced globally in the telecommunications industry. The objective of these exercises is to build a classification model to measure the propensity of an active good customer to churn. This will enable the service provider to take active steps to retain the profitable customers before churn occurs. The data set that is used in these exercises has rare target events, and because of this, the data set is oversampled. This is because you tend to get better models when they are built on a data set that is more balanced with respect to the levels of the target variable. Your data is oversampled; the probabilities in the population are 4% churners and 96% non-churners. Enterprise Miner enables you to enter this information before or after the modeling is complete. You have chosen to enter it at the beginning of the process flow in the diagram, and you enter it by going to the data node and then in the Property panel selecting the **Decisions** button.

* 1. Select **New Project**.
  2. Name the project **Churn** andspecify the path to the data as **your H or M drive.**
  3. Create a diagram named **Churn\_Model**.
  4. Add the data set **Churn\_RPM** as source data for the project. The metadata data for this data set is presented below.



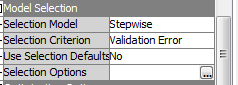
* 1. Explore **Churn\_RPM**.
  2. In the data node properties panel, select **Decisions**,select **Build**, and enter the prior probabilities.



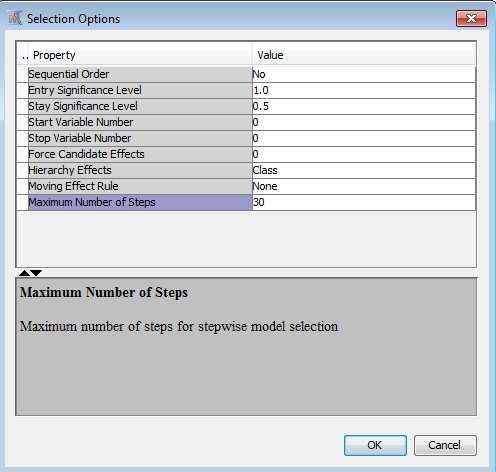
* 1. Now select the **Decisions** tab and select . Selecting the Decision Weights tab shows that the inverse weights are 25 for Decision1 and 1.0416667 for Decision2, and this changes the cutoff of probabilities to be .04. This option is chosen because the probability of churn is small.
  2. Do some additional data exploration by using some of the nodes from the **Explore** tab of **SEMMA**.
  3. Add **Partition** and **Decision Tree** nodes to the Process Flow.
     1. In the properties panel for the **Partition** node set Training to **.70**, Validation to **.30**, and Test to **0**.
     2. Build several trees and compare their results. For each tree, determine the important variables.
     3. How many leaves defined the maximum tree for each tree model you built?
  4. Add **Transformation**, **Impute**, and **Regression** nodes to the process flow.
     1. Determine what variables need transforming and select **Max. Normal** as the method.
     2. In the **Impute** node, select **Count** as the imputation method for class variables and **Tree** for interval.

Is imputation necessary for interval variables? Create a unique indicator variable for missing Variables.

* + 1. In the property panel for the Regression node, make the changes as shown below.



* 1. Now select the ellipsis button next to **Selection Options** and make the changes shown below.



* 1. Runthe **Regression** nodes and compare results with the **Tree** nodes results.
  2. Score current good customers.
     1. Score the data set **Churn\_RPM\_Score** using your selected model; table role must be **Score**.  
        This data set has fifty customers that have been identified for scoring.
     2. How many customers did your model classify as churners?

1. **Setting Up a Project for the Annuity Data Set**

A bank seeks to increase sales of a variable annuity product. To do this, the bank will send product offers to existing banking customers. However, to maximize profits, the bank wants to be selective about whom it targets. This selectivity will be achieved by constructing a predictive model. The data set name is **Annuity**.

To achieve the bank’s analytic objective, an analysis data set was assembled. The data set contains 10,619 records and 48 variables, assembled from several source tables within the bank’s data warehouse. The source tables include the customer master table, the transaction detail table, the product detail table, and a third-party demographic overlay table. The variables describe each customer’s demographics and usage of other banking products prior to acquisition of the variable annuity. Two of the variables are nominally scaled; the remainder are binary or interval.

About half of the variables have some missing values. Many of the variables, especially those relating to monetary amounts, have an extremely large range and highly skewed distribution.

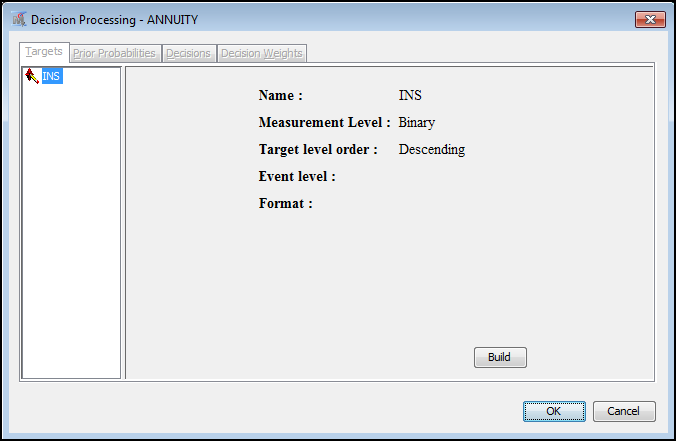
The **BRANCH** variable, a nominal input with 19 distinct levels, indicates the branch in which the customer’s initial account was opened. The **RES** variable, a nominal input with three distinct levels, classifies the customer’s primary residence as rural, suburban, or urban.

The target variable for this analysis, **INS**, indicates acquisition of the variable annuity over a fixed period of time. While overall acquisition rate is about 2%, the acquisition rate in the raw analysis data is more than 34%. This reflects the separate sampling used to generate the raw data.

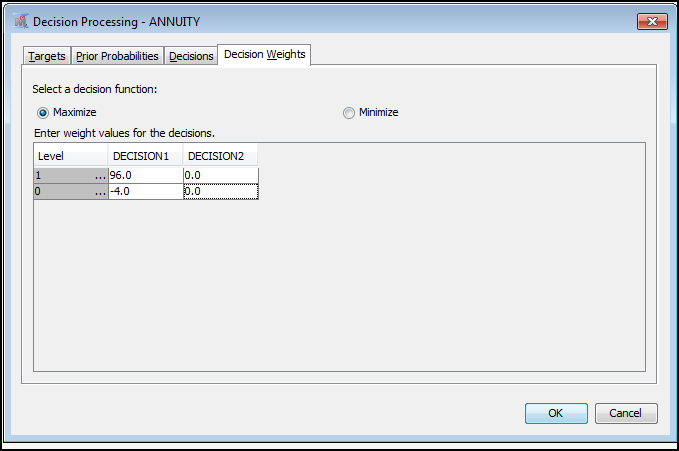
The bank expects to realize an average short-term revenue of about $100 from each customer who purchases the annuity product. It is expected to cost the bank about $4 per solicitation (which involves an initial mail solicitation with a telephone follow-up) to carry out the campaign.

**Decision Processing**

The information given above enables you to create a profit loss matrix and new assessment statistics for this exercise. This window also enables you to adjust for oversampling. When the data set **Annuity** is brought into a diagram, select the ellipsis button next to **Decisions**.



* 1. Select the button **Build**. The tabs are now active.
  2. Select **Prior Probabilities** ⇨ **Yes** and enter **.02** and **.98** for **Adjusted Prior**.
  3. Select the **Decision Weights** tab and create the Profit Loss Matrix shown below.



If a person takes out an annuity, the average profit is $96, and if a person does not respond to the solicitation, the loss is $4. This changes the cutoff for probabilities to .04 rather than .5.

* 1. Build your project and open the Decision Processing window by following the steps above.
  2. Explore the data and build predictive models.
  3. Select the “best model” based on your assessment statistic and the validation data set.

1. **Building Predictive Models for Bank\_1\_17\_14\_new Data Set**

This is the data set that was created in Chapter 2 for predictive modeling.

Build predictive models and find the “best model”.

1. **Using RPM to Create Models for the Retention Data Set**
   1. Create the data set **Retention** for the **RetenRPM** project and define the metadata as you did in the examples.
   2. Run the diagram with the **Retention** data set and compare results obtained with those in the examples.
   3. Go to Enterprise Guide and open the project **RetenRPM** and create the advanced model. Examine the results from this diagram in Enterprise Miner.