Introduction to Predictive Modeling

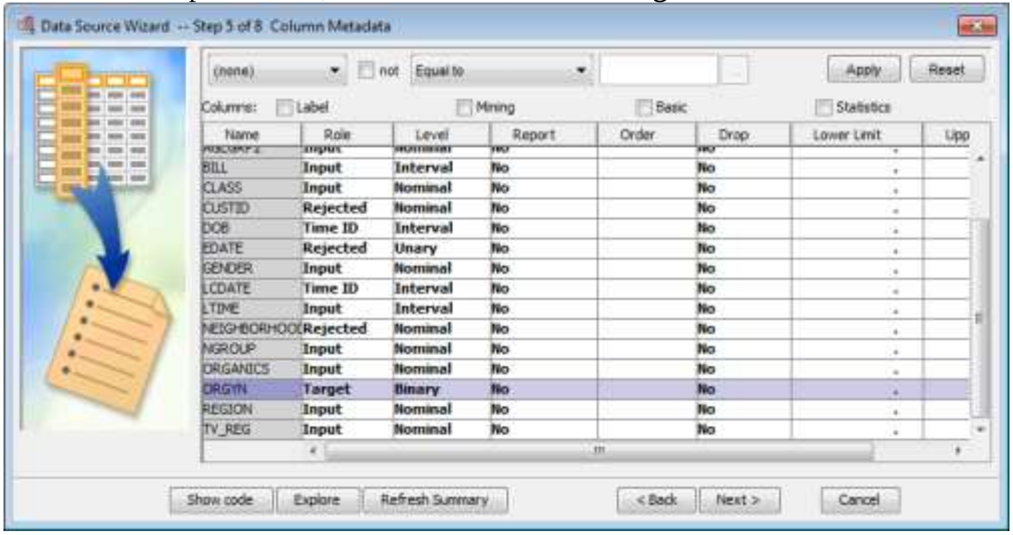
Define the data source for the Organic Purchase Analysis project.

Your Task

Define **AAEM.ORGANICS** as a data source in the **eLearnEM** project and adjust the column metadata as described in the following steps.

1. Specify **AAEM.Organics** as the data source.
   Allow SAS Enterprise Miner to automatically determine the variable roles and measurement levels for you. Use the default settings.
   - In the project panel, right-click **Data Sources** and select **Create Data Source**.
   - On the Metadata Source step of the Data Source Wizard, be sure that **SAS Table** is selected, and click **Next**.
   - In the Select a SAS Table step, click **Browse**.
   - Double-click **AAEM.**
   - Select **ORGANICS** and click **OK**.
   - In the Select a SAS Table step, click **Next**.
   - In the Table Information step, click **Next**.
   - In the Metadata Advisor Options step, select **Advanced**, then click **Next**.

2. Specify **ORGYN** as the target.
   - In the **Column Metadata** table, click in the **Role** column for **ORGYN**.
   - In the drop-down list, scroll down and select **Target**.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upp
BILL	Input	Interval	No		No	.	
CLASS	Input	Nominal	No		No	.	
CUSTID	Rejected	Nominal	No		No	.	
DOB	Time ID	Interval	No		No	.	
EDATE	Rejected	Unary	No		No	.	
GENDER	Input	Nominal	No		No	.	
LCDATE	Time ID	Interval	No		No	.	
LTIME	Input	Interval	No		No	.	
NEIGHBORHOOI	Rejected	Nominal	No		No	.	
NGROUP	Input	Nominal	No		No	.	
ORGANICS	Input	Nominal	No		No	.	
ORGYN	Target	Binary	No		No	.	
REGION	Input	Nominal	No		No	.	
TV_REG	Input	Nominal	No		No	.	

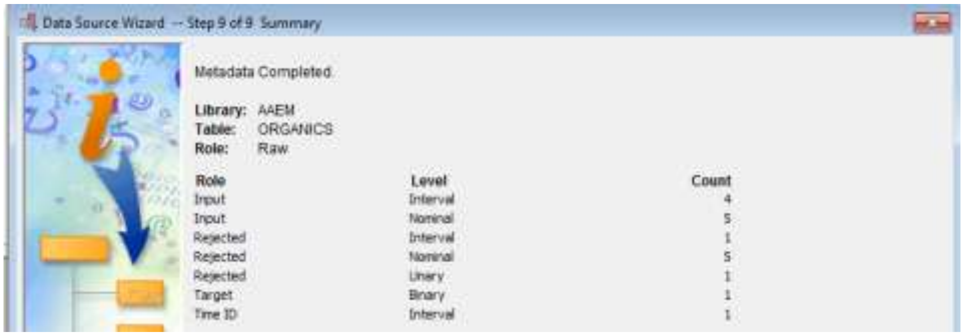
3. Set the model role for **AGEGRP1** and **AGEGRP2** to **Rejected**.
   The variables **AGE**, **AGEGRP1**, and **AGEGRP2** are all different measurements for the same information. Presume that, based on previous experience, you know that **AGE** should be used for this type of modeling.
   - Click in the **Role** column for **AGEGRP1**.
   - In the drop-down list, scroll down and select **Rejected**.
   - Click in the **Role** column for **AGEGRP2**.
   - In the drop-down list, scroll down and select **Rejected**.

4. Set the model role for **ORGANICS** to **Rejected**.
   ORGANICS contains information that would not be known at the time that you are developing a model to predict the purchase of organic products.
   - Click in the **Role** column for **ORGANICS**.
   - In the drop-down list, scroll down and select **Rejected**.

5. Notice that the numeric variable **LCDATE** is automatically rejected. Why do you think this is the case?
   LCDATE and **LTIME** essentially measure the same thing. Presume that **LTIME** is sufficient for building your predictive models.

   Notes: LCDATE was not automatically rejected. I manually rejected.

   Answer: LCDATE is rejected because it contains more than 50 missing values.
6. Notice that the character variable **NEIGHBORHOOD** is also automatically rejected. Why do you think this is the case?
   NGROUP contains collapsed levels of **NEIGHBORHOOD**. Presume that **NGROUP** is sufficient for building your predictive models.
   Answer: NEIGHBORHOOD is rejected because it has a class count that is greater than 20.
7. Specify the decision configuration and data source attributes. Accept the default settings.
   - Click **Next**.
   - In the Decision Configuration step, verify that **No** is selected and click **Next**.
   - In the Create Sample step, verify that **No** is selected and click **Next**.
   - In the Data Source Attributes step, verify that the role is set to **Raw** and then click **Next**.
   - Click **Finish**.



Metadata Completed

Library: AAEM
Table: ORGANICS
Role: Raw

Role	Level	Count
Input	Interval	4
Input	Nominal	5
Rejected	Interval	1
Rejected	Nominal	5
Rejected	Unary	1
Target	Binary	1
Time ID	Interval	1

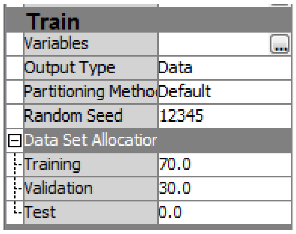
Create a diagram and partition the input data.

Your Task
Create a diagram named **Organics** and partition the input data.

1. Create a diagram named **Organics** and add the **ORGANICS** data source to the diagram workspace.
   - In the project panel, right-click **Diagrams** and select **Create Diagram**.
   - In the **Name** box, type Organics, and then click **OK**.
   - In the project panel, expand **Data Sources**.
   - Add an **ORGANICS** node to the diagram.

2. Add a **Data Partition** node to the diagram and connect it to the **Data Source** node. Assign 70% of the data for training and 30% for validation.
   - From the **Sample** tab toolbar, add a **Data Partition** node to the diagram.

- In the diagram, connect the **ORGANICS** node to the **Data Partition** node.
- In the diagram, click the **Data Partition** node.
- In the properties panel, type 70 in the **Value** column for the **Training** property. Then press **ENTER**.
- In the **Value** column for the **Test** property, type 0. Then press **ENTER.**



Train	
Variables	...
Output Type	Data
Partitioning Metho	Default
Random Seed	12345
⊟ Data Set Allocatior	
Training	70.0
Validation	30.0
Test	0.0

3. Run the path and view the results.
   When you finish viewing the results, close the **Results** window.
   - Right-click the **Data Partition** node and select **Run**.
   - In the Confirmation window, click **Yes**.
   - When the Run Status window opens, click **Results**.
   - Notice that the Output window displays a variable summary and summary statistics for targets in the original data and in the partition data sets.
   - In the Output window, scroll down to view any additional output, then close the Results window.

```
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*

Variable Summary

            Measurement    Frequency
Role           Level         Count

INPUT        INTERVAL          4
INPUT        NOMINAL           5
REJECTED     INTERVAL          1
REJECTED     NOMINAL           5
REJECTED     UNARY             1
TARGET       BINARY            1
TIMEID       INTERVAL          1


Partition Summary

                                     Number of
Type             Data Set          Observations

DATA        EMWS8.Ids_DATA             22223
TRAIN       EMWS8.Part_TRAIN           15557
VALIDATE    EMWS8.Part_VALIDATE         6666
*------------------------------------------------------------*

Summary Statistics for Class Targets

Data=DATA

            Numeric    Formatted    Frequency
Variable     Value       Value        Count      Percent           Label

 ORGYN         0           0          16718      75.2284    Organics Purchased?
 ORGYN         1           1           5505      24.7716    Organics Purchased?


Data=TRAIN

            Numeric    Formatted    Frequency
Variable     Value       Value        Count      Percent           Label

 ORGYN         0           0          11703      75.2266    Organics Purchased?
 ORGYN         1           1           3854      24.7734    Organics Purchased?
```

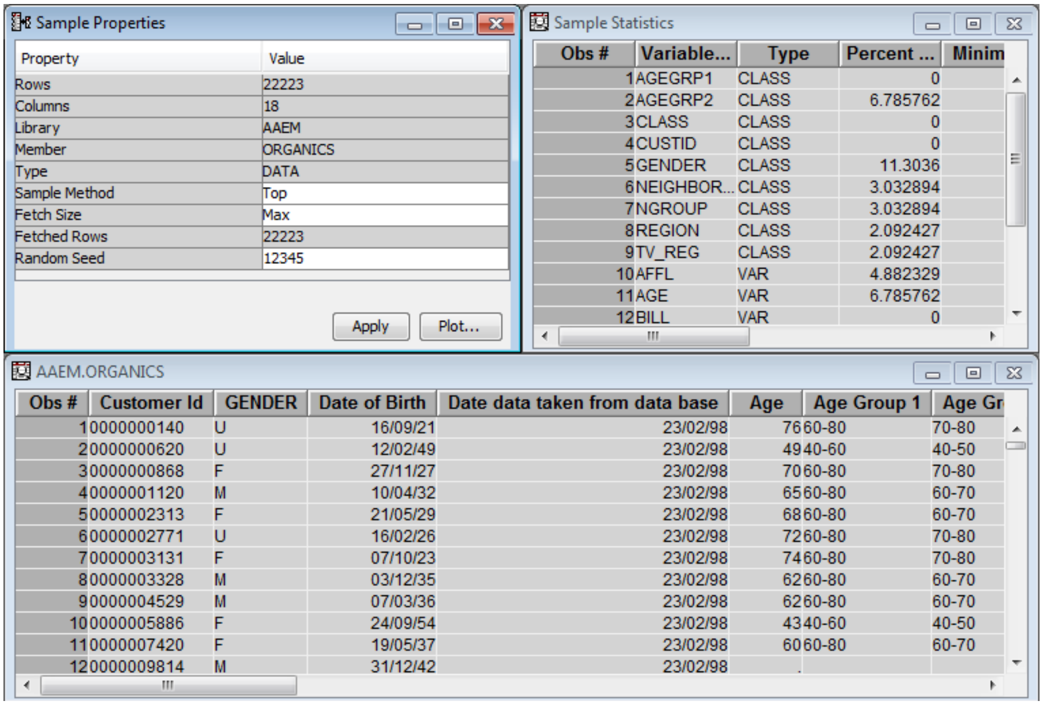
```
Data=VALIDATE

             Numeric     Formatted     Frequency
Variable      Value        Value         Count        Percent            Label

 ORGYN          0            0            5015        75.2325     Organics Purchased?
 ORGYN          1            1            1651        24.7675     Organics Purchased?
```

Regression Models

Your Task

Suppose you want to determine whether missing value imputation is needed as preparation for regression on the **ORGANICS** data source. You explore the **ORGANICS** data source and decide to impute missing values and create indicator variables. Then you perform a regression analysis on imputed values.

1. Explore the **ORGANICS** data source.
   - In the project panel, right-click the **ORGANICS** data source and select **Explore.**
   - If you see a Large Data Constraint window, click **OK**.



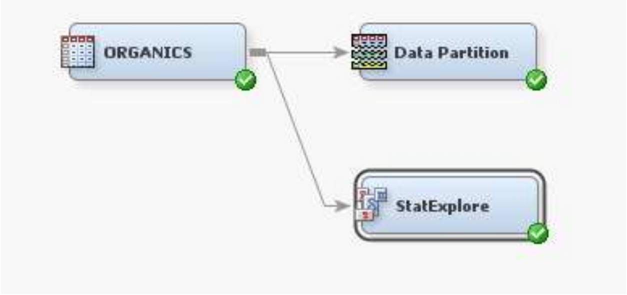
Property	Value
Rows	22223
Columns	18
Library	AAEM
Member	ORGANICS
Type	DATA
Sample Method	Top
Fetch Size	Max
Fetched Rows	22223
Random Seed	12345

Obs #	Variable...	Type	Percent ...	Minim
1	AGEGRP1	CLASS	0	
2	AGEGRP2	CLASS	6.785762	
3	CLASS	CLASS	0	
4	CUSTID	CLASS	0	
5	GENDER	CLASS	11.3036	
6	NEIGHBOR...	CLASS	3.032894	
7	NGROUP	CLASS	3.032894	
8	REGION	CLASS	2.092427	
9	TV_REG	CLASS	2.092427	
10	AFFL	VAR	4.882329	
11	AGE	VAR	6.785762	
12	BILL	VAR	0	

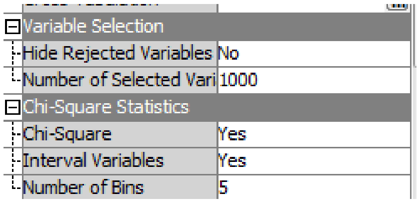
Obs #	Customer Id	GENDER	Date of Birth	Date data taken from data base	Age	Age Group 1	Age Gr
1	0000000140	U	16/09/21	23/02/98	76	60-80	70-80
2	0000000620	U	12/02/49	23/02/98	49	40-60	40-50
3	0000000868	F	27/11/27	23/02/98	70	60-80	70-80
4	0000001120	M	10/04/32	23/02/98	65	60-80	60-70
5	0000002313	F	21/05/29	23/02/98	68	60-80	60-70
6	0000002771	U	16/02/26	23/02/98	72	60-80	70-80
7	0000003131	F	07/10/23	23/02/98	74	60-80	70-80
8	0000003328	M	03/12/35	23/02/98	62	60-80	60-70
9	0000004529	M	07/03/36	23/02/98	62	60-80	60-70
10	0000005886	F	24/09/54	23/02/98	43	40-60	40-50
11	0000007420	F	19/05/37	23/02/98	60	60-80	60-70
12	0000009814	M	31/12/42	23/02/98	.		

   - Examine the **AAEM.ORGANICS** data table. Maximize the window. Because you need to scroll down many times to see missing values for all observations, it is best to use another method to check for missing values.
   - Close the Explore window.
2. Open the **Organics** diagram.
3. Use the **StatExplore** node to more easily examine missing data values. Change the value for the property for **Hide Rejected Variables** to *No* and the value for the property **Interval Variables** to *Yes*.
   - Click the **Explore** tab. Add a **StatExplore** node to the diagram.

- Connect the **ORGANICS** node to the **StatExplore** node.

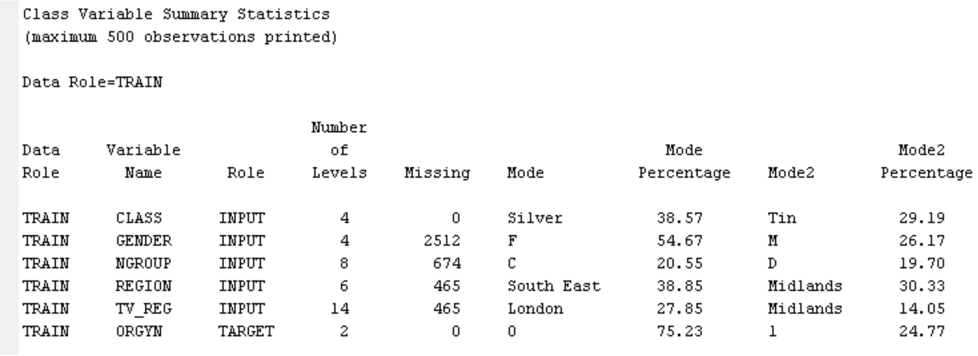


- Select the **StatExplore** node in the diagram and examine the properties panel.
- For the **Hide Rejected Variables** property, Select **No**. For the **Interval Variables** property, Select **Yes**.

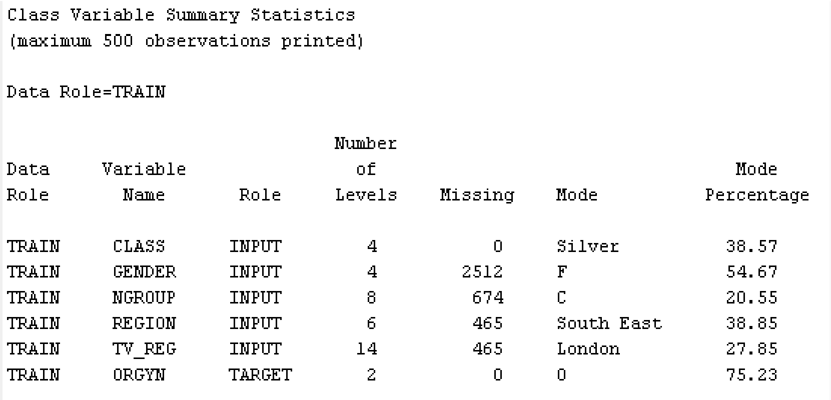


```
⊟Variable Selection
 ├Hide Rejected Variables  No
 └Number of Selected Vari  1000
⊟Chi-Square Statistics
 ├Chi-Square               Yes
 ├Interval Variables       Yes
 └Number of Bins           5
```

- Click **Run**.
- In the Confirmation window, click **Yes**.
- In the Run Status window, click **Results**.
- Maximize the Output window. Scroll down to **Class Variable Summary Statistics**. Notice that the variable **GENDER** has a relatively large number of missing values.

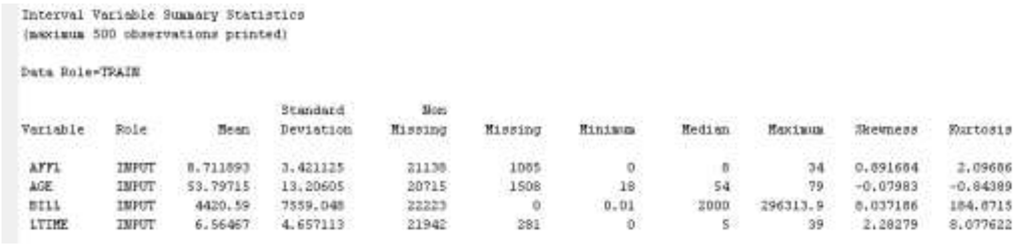


```
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                              Number
Data     Variable               of                                   Mode                     Mode2
Role       Name      Role     Levels    Missing   Mode            Percentage    Mode2       Percentage

TRAIN     CLASS      INPUT       4          0     Silver             38.57      Tin            29.19
TRAIN     GENDER     INPUT       4       2512     F                  54.67      M              26.17
TRAIN     NGROUP     INPUT       8        674     C                  20.55      D              19.70
TRAIN     REGION     INPUT       6        465     South East         38.85      Midlands       30.33
TRAIN     TV_REG     INPUT      14        465     London             27.85      Midlands       14.05
TRAIN     ORGYN      TARGET      2          0     0                  75.23      1              24.77
```



```
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                              Number
Data     Variable               of                                   Mode
Role       Name      Role     Levels    Missing   Mode            Percentage

TRAIN     CLASS      INPUT       4          0     Silver             38.57
TRAIN     GENDER     INPUT       4       2512     F                  54.67
TRAIN     NGROUP     INPUT       8        674     C                  20.55
TRAIN     REGION     INPUT       6        465     South East         38.85
TRAIN     TV_REG     INPUT      14        465     London             27.85
TRAIN     ORGYN      TARGET      2          0     0                  75.23
```

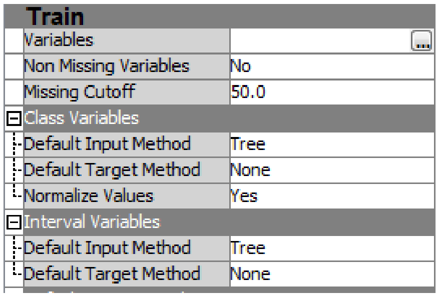
- Scroll down to **Interval Variable Summary Statistics**. Notice that the variables **AFFL** and **AGE** have over 1000 missing values each.



```
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                  Standard        Non
Variable    Role        Mean     Deviation    Missing    Missing    Minimum     Median     Maximum    Skewness    Kurtosis

 AFFL       INPUT   8.711893     3.421125      21138       1085          0          8          34    0.891684     2.09686
 AGE        INPUT   53.79715     13.20605      20715       1508         18         54          79    -0.07983    -0.84389
 BILL       INPUT    4420.59     7559.048      22223          0       0.01       2000    296313.9    8.037186    184.8715
 LTIME      INPUT    6.56467     4.657113      21942        281          0          5          39     2.28279    8.077622
```

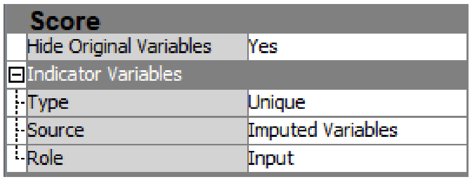
- Close the Results window.

4. Add an **Impute** node to the diagram and connect it to the **Data Partition** node. Change the default input method to *Tree* for both class and interval variables. *Tree* is used as an estimation method for imputing missing values.
   - Click the **Modify** tab. Add an **Impute** node to the diagram.
   - Connect the **Data Partition** node to the **Impute** node.
   - Select the **Impute** node.
   - Under the heading **Class Variables**, for the **Default Input Method** property, select **Tree**.
   - Under the heading **Interval Variables,** for the **Default Input Method** property, select **Tree**.



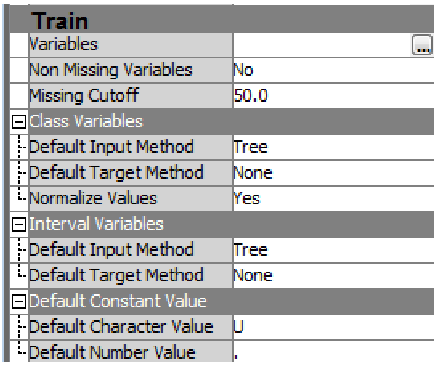
Train	
Variables	...
Non Missing Variables	No
Missing Cutoff	50.0
⊟Class Variables	
Default Input Method	Tree
Default Target Method	None
Normalize Values	Yes
⊟Interval Variables	
Default Input Method	Tree
Default Target Method	None

5. Create missing value indicator variables that can serve as new inputs that are unique. Change the property **Indicator Variables** to *Unique* and the property **Indicator Variable Role** to *Input*.



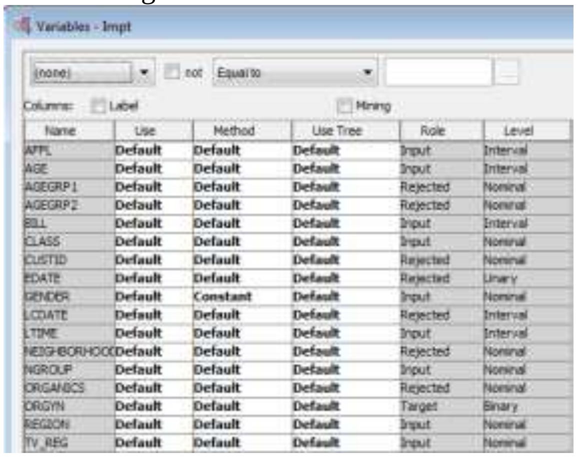
Score	
Hide Original Variables	Yes
⊟Indicator Variables	
Type	Unique
Source	Imputed Variables
Role	Input

6. Replace missing values for **GENDER** with *U* for unknown.
   - Scroll up to **Default Constant Value**. Under this heading, click in the **Value** column for the **Default Character Value** property and type **U**



Train	
Variables	...
Non Missing Variables	No
Missing Cutoff	50.0
⊟ Class Variables	
Default Input Method	Tree
Default Target Method	None
Normalize Values	Yes
⊟ Interval Variables	
Default Input Method	Tree
Default Target Method	None
⊟ Default Constant Value	
Default Character Value	U
Default Number Value	.

7. Use the Variables window to change the method for the variable **GENDER** to *Constant*.



Variables - Impt

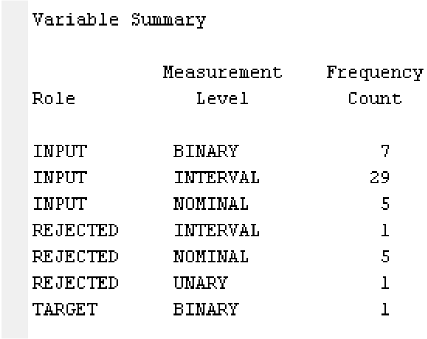
(none) ▾ □ not Equal to ▾

Columns: □ Label □ Mining

Name	Use	Method	Use Tree	Role	Level
APPL	Default	Default	Default	Input	Interval
AGE	Default	Default	Default	Input	Interval
AGEGRP1	Default	Default	Default	Rejected	Nominal
AGEGRP2	Default	Default	Default	Rejected	Nominal
BILL	Default	Default	Default	Input	Interval
CLASS	Default	Default	Default	Input	Nominal
CUSTID	Default	Default	Default	Rejected	Nominal
EDATE	Default	Default	Default	Rejected	Unary
GENDER	Default	Constant	Default	Input	Nominal
LCDATE	Default	Default	Default	Rejected	Interval
LTIME	Default	Default	Default	Input	Interval
NEIGHBORHOOD	Default	Default	Default	Rejected	Nominal
NGROUP	Default	Default	Default	Input	Nominal
ORGANICS	Default	Default	Default	Rejected	Nominal
ORGYN	Default	Default	Default	Target	Binary
REGION	Default	Default	Default	Input	Nominal
TV_REG	Default	Default	Default	Input	Nominal

8. Add a **Regression** node to the diagram and connect it to the **Impute** node.
9. Run the **Regression** node and display the results. In the output, review the **Variable Summary** information. How many inputs predict target variables?

- Select the **Regression** node and click **Run**.
- In the Confirmation window, click **Yes**.
- In the Run Status window, click **Results**.
- Maximize the Output window. Review the **Variable Summary** information at or near the top of the window. Note that 16 inputs predict target variables. (I got different outcome)
- Close the Results window.



```
Variable Summary

             Measurement    Frequency
Role            Level         Count

INPUT         BINARY             7
INPUT         INTERVAL          29
INPUT         NOMINAL            5
REJECTED      INTERVAL           1
REJECTED      NOMINAL            5
REJECTED      UNARY              1
TARGET        BINARY             1
```