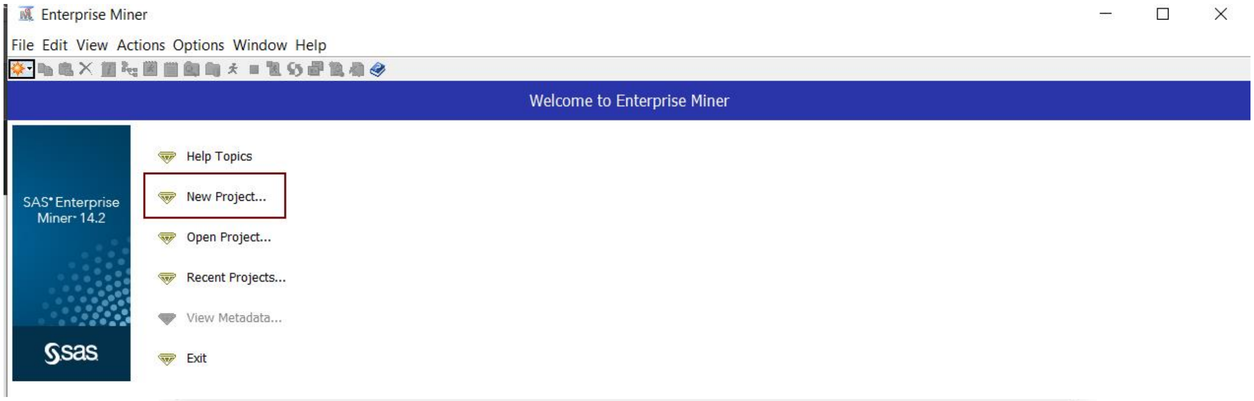
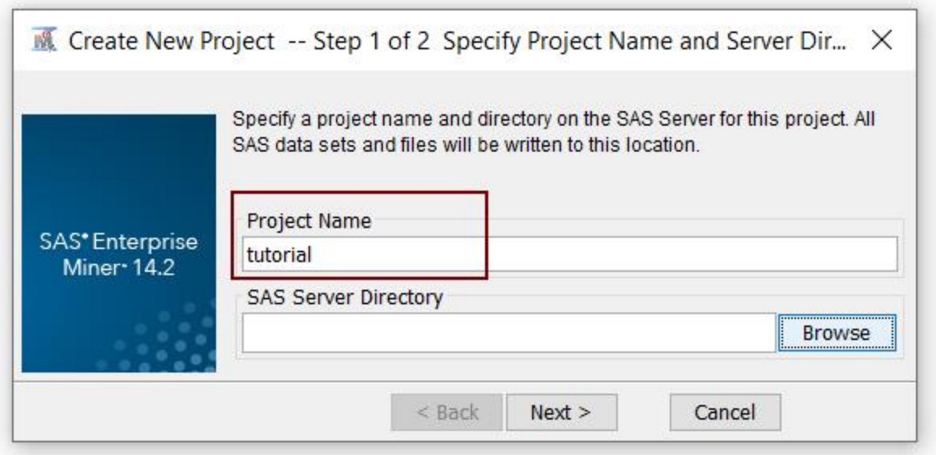
Contents

WHERE TO START

1) OPEN A NEW PROJECT:

OPEN A NEW PROJECT → PROJECT NAME (TYPE IN THE NAME) → SAS SERVER DIRECTORY →BROWSE → FIND THE LOCATION TO SAVE YOUR PROJECT → OPEN → NEXT → FINISH

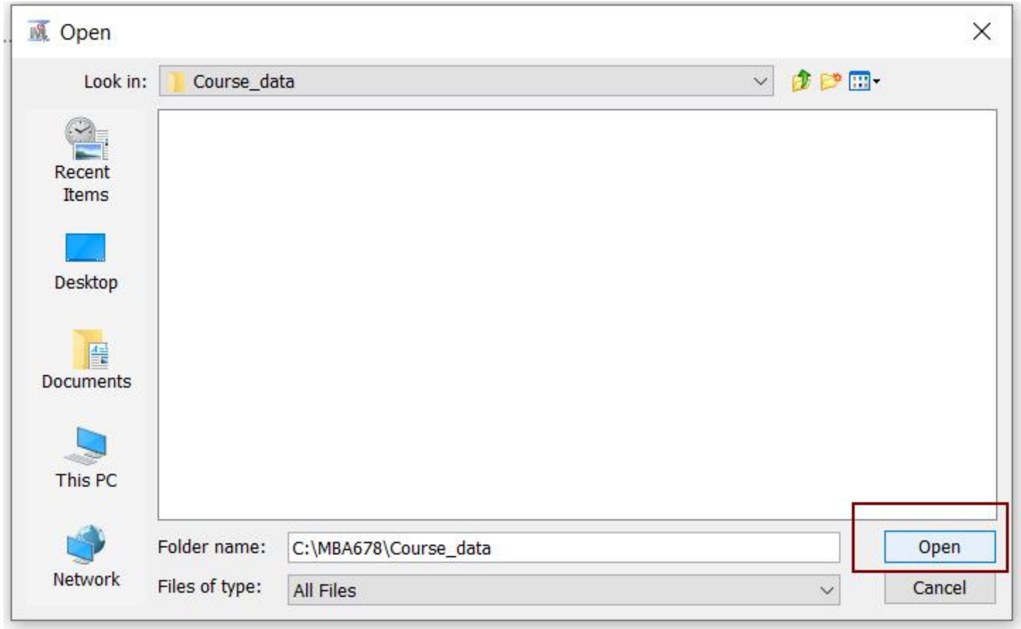

Enterprise Miner
File Edit View Actions Options Window Help
Welcome to Enterprise Miner
SAS Enterprise Miner 14.2
Help Topics
New Project...
Open Project...
Recent Projects...
View Metadata...
Exit


Create New Project -- Step 1 of 2 Specify Project Name and Server Dir...
SAS Enterprise Miner 14.2
< Back
Next >
Cancel
Browse

Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

Project Name: tutorial

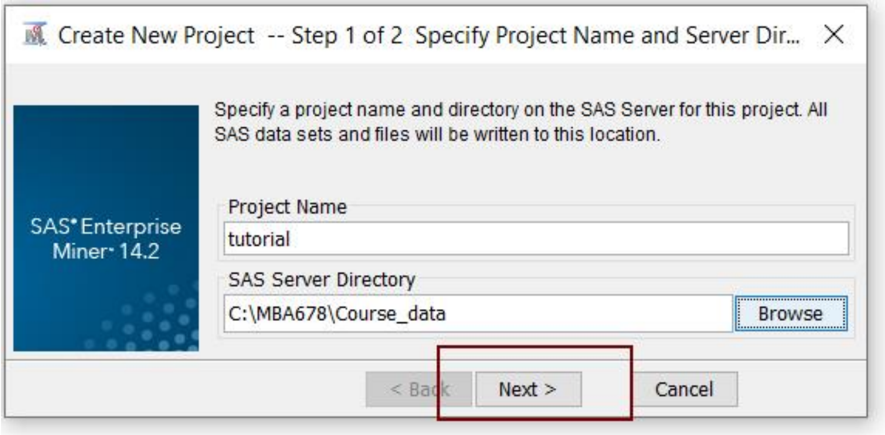
SAS Server Directory:



Look in: Course_data

Folder name: C:\MBA678\Course_data

Files of type: All Files

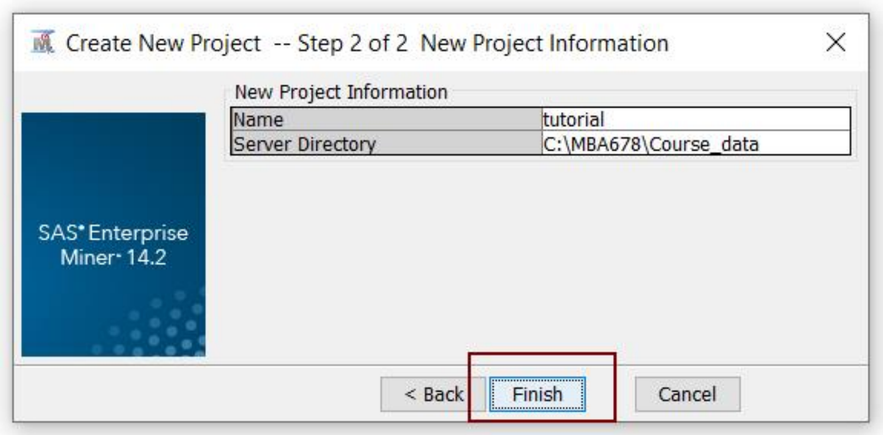


Create New Project -- Step 1 of 2 Specify Project Name and Server Dir...

Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

Project Name: tutorial

SAS Server Directory: C:\MBA678\Course_data



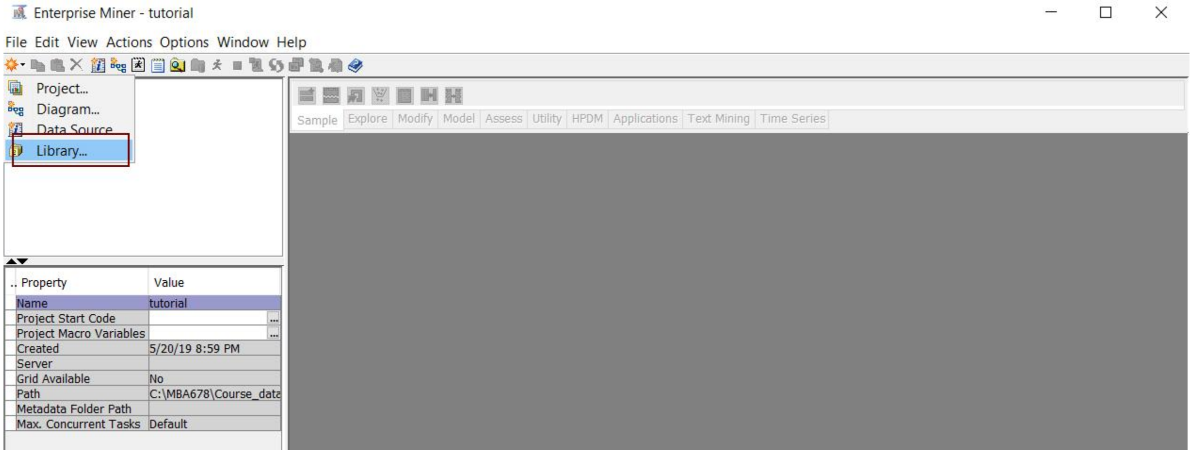
Create New Project -- Step 2 of 2 New Project Information

New Project Information

Name	tutorial
Server Directory	C:\MBA678\Course_data

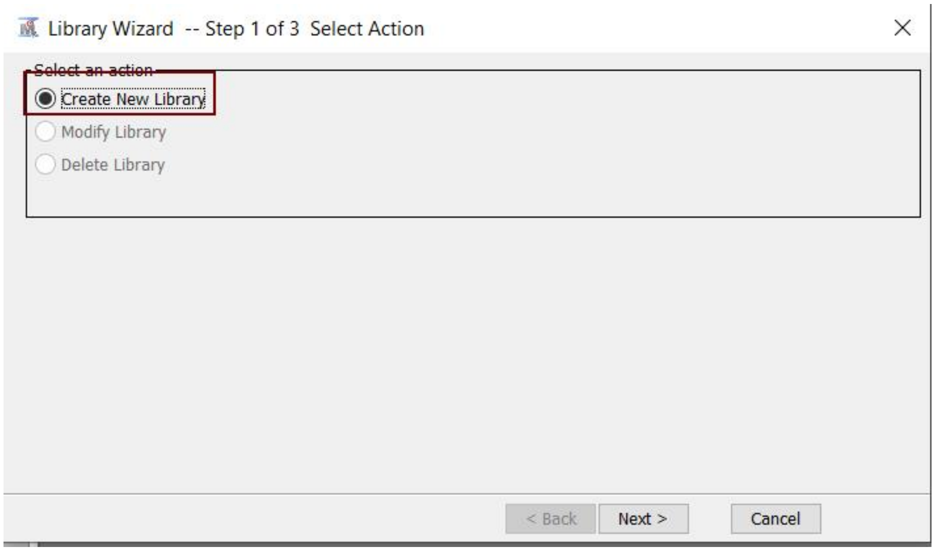
2) CREATE A NEW LIBRARY:

LIBRARY → CREATE NEW LIBRARY → NEXT → NAME (TYPE IN THE LIBRARY NAME) → PATH: BROWSE (CHOOSE A LOCATION FOR YOUR NEW LIBRARY) → OPEN → NEXT → FINISH



Enterprise Miner - tutorial

Property	Value
Name	tutorial
Project Start Code	
Project Macro Variables	
Created	5/20/19 8:59 PM
Server	
Grid Available	No
Path	C:\MBA678\Course_data
Metadata Folder Path	
Max. Concurrent Tasks	Default



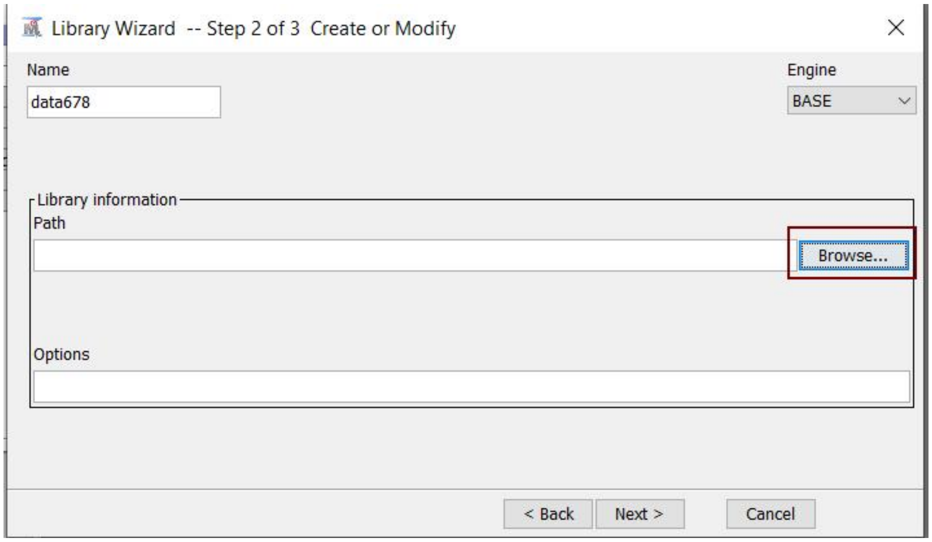
Library Wizard -- Step 1 of 3 Select Action

Select an action

● Create New Library

○ Modify Library

○ Delete Library



Library Wizard -- Step 2 of 3 Create or Modify

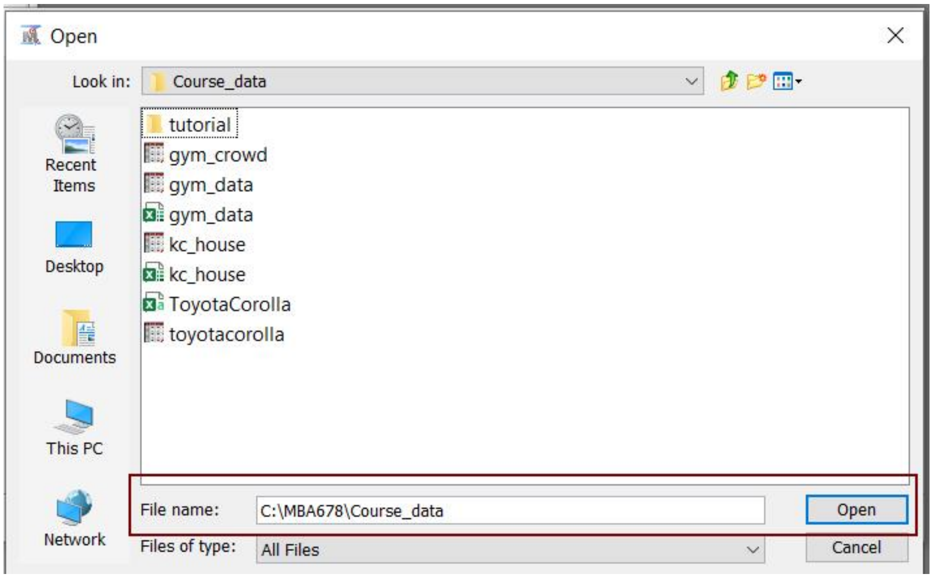
Name: data678

Engine: BASE

Library information

Path:

Options:



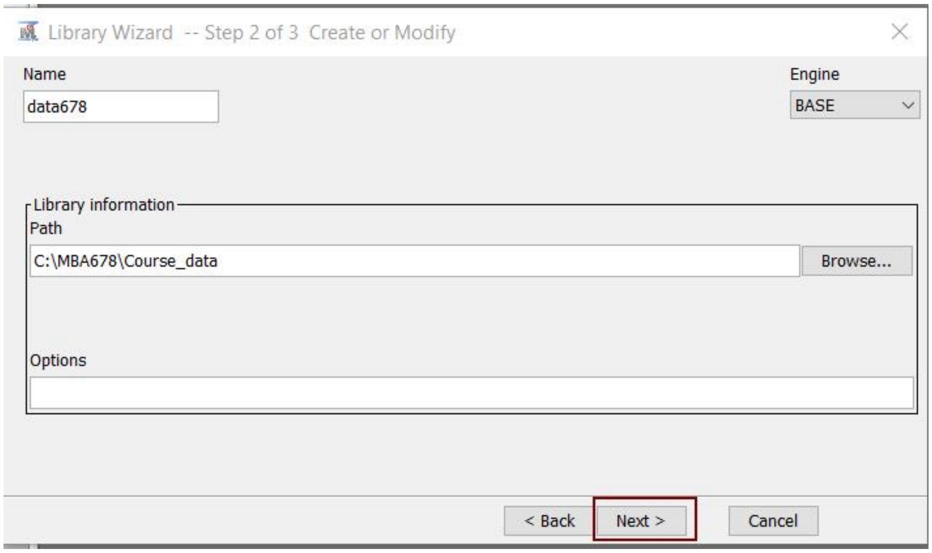
Open

Look in: Course_data

- tutorial
- gym_crowd
- gym_data
- gym_data
- kc_house
- kc_house
- ToyotaCorolla
- toyotacorolla

File name: C:\MBA678\Course_data

Files of type: All Files



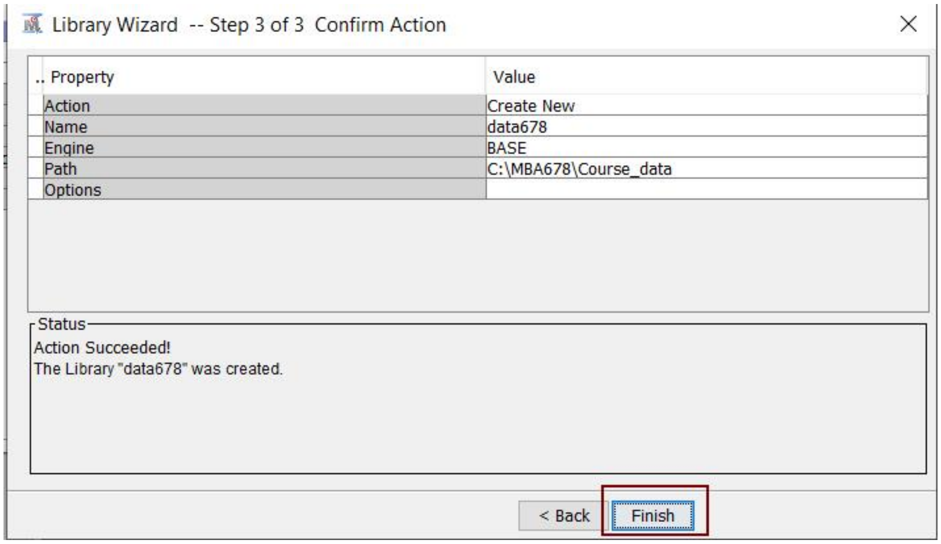
Name: data678

Engine: BASE

Library information

Path: C:\MBA678\Course_data

Options



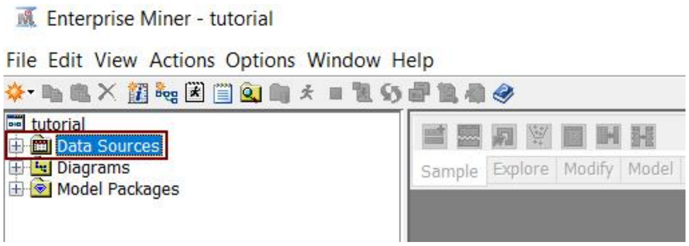
.. Property	Value
Action	Create New
Name	data678
Engine	BASE
Path	C:\MBA678\Course_data
Options	

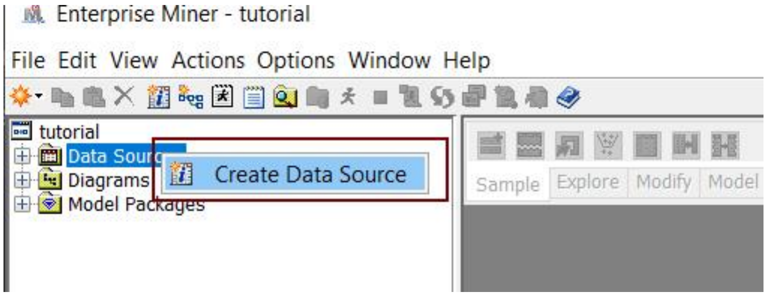
Status

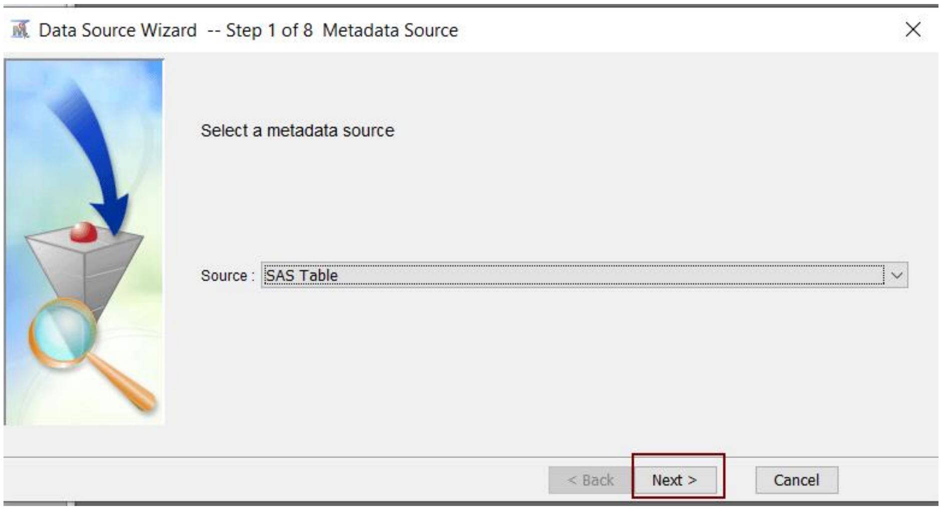
Action Succeeded!
The Library "data678" was created.

3) CREATE A DATA SOURCE:

right-click on **DATA SOURCES → left-click on CREATE DATA SOURCE → SAS TABLE (LEAVE AT DEFAULT) → NEXT → TABLE: BROWSE → CHOOSE THE LIBRARY(DOUBLE CLICK) → CHOOSE THE FILE → OK → NEXT → STEP 3: LEAVE AT DEFAULT → NEXT → METADATA ADVISOR OPTIONS: BASIC (LEAVE AT DEFAULT) → NEXT → STEP 5: ASSIGN VARIABLE ROLES (!!!SET THE BINARY VARIABLE YOU ARE TRYING TO PREDICT TO TARGET AND ITS LEVEL TO BINARY) → CONTINUE CHANGING THE VARIABLE ROLES AND LEVELS IF NEEDED → NEXT → STEP 6 (LEAVE AT DEFAULT): NEXT → STEP 7 (LEAVE AT DEFAULT): NEXT → STEP 8 (LEAVE AT DEFAULT): NEXT → FINISH**

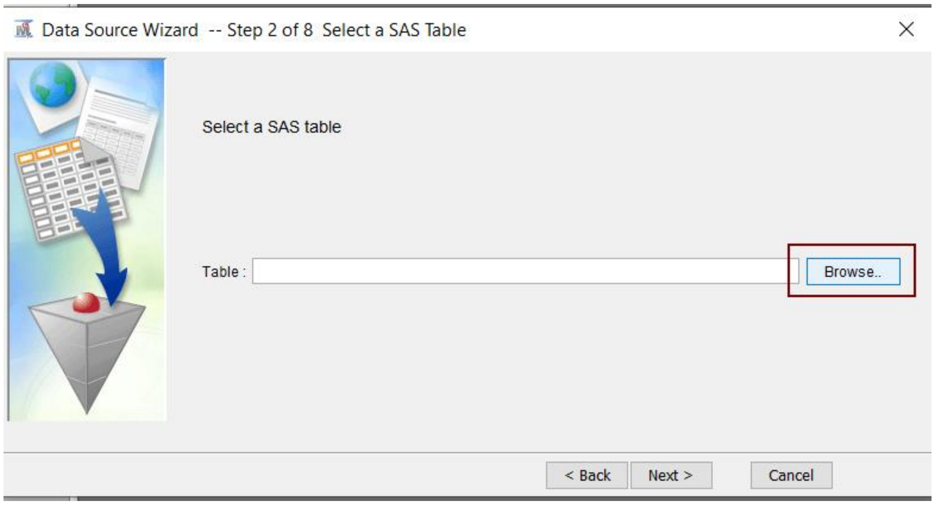

Enterprise Miner - tutorial
File Edit View Actions Options Window Help
tutorial
Data Sources
Diagrams
Model Packages
Sample
Explore
Modify
Model





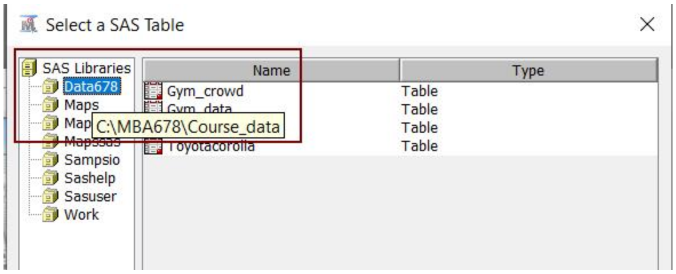
Select a metadata source

Source : SAS Table

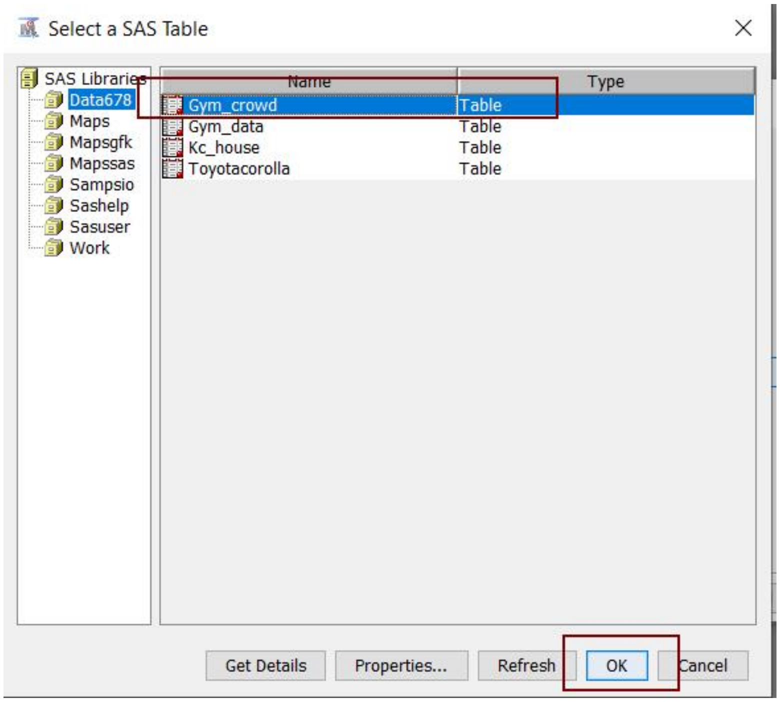


Select a SAS table

Table :



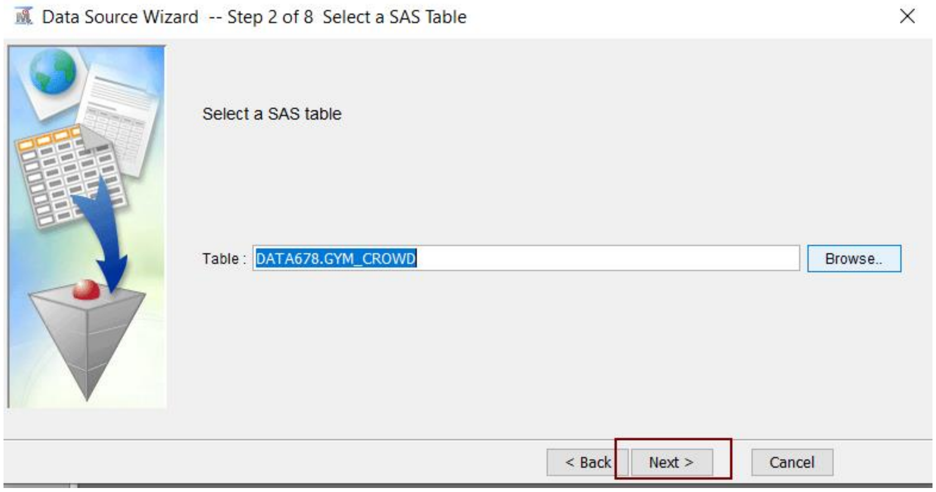
Name	Type
Gym_crowd	Table
Gym_data	Table
	Table
Toyotacorolla	Table



Select a SAS Table

SAS Libraries
- Data678
- Maps
- Mapsgfk
- Mapssas
- Sampsio
- Sashelp
- Sasuser
- Work

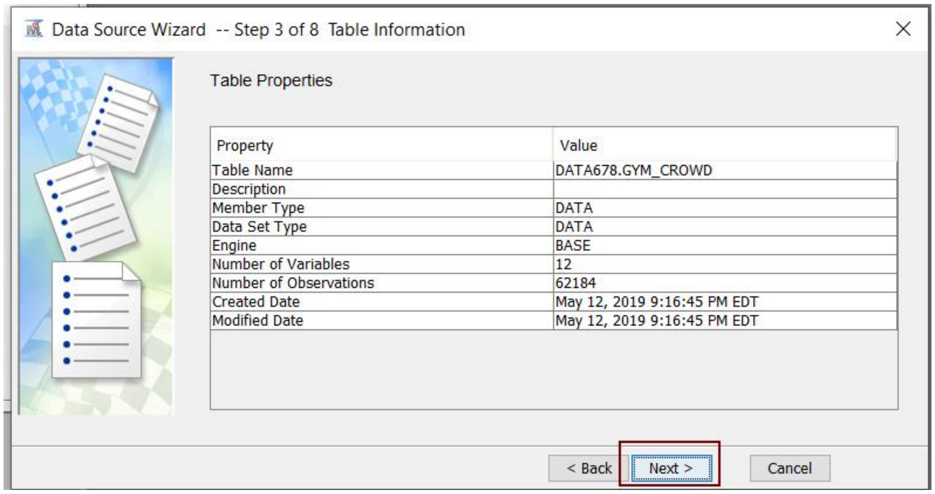
Name	Type
Gym_crowd	Table
Gym_data	Table
Kc_house	Table
Toyotacorolla	Table



Data Source Wizard -- Step 2 of 8 Select a SAS Table

Select a SAS table

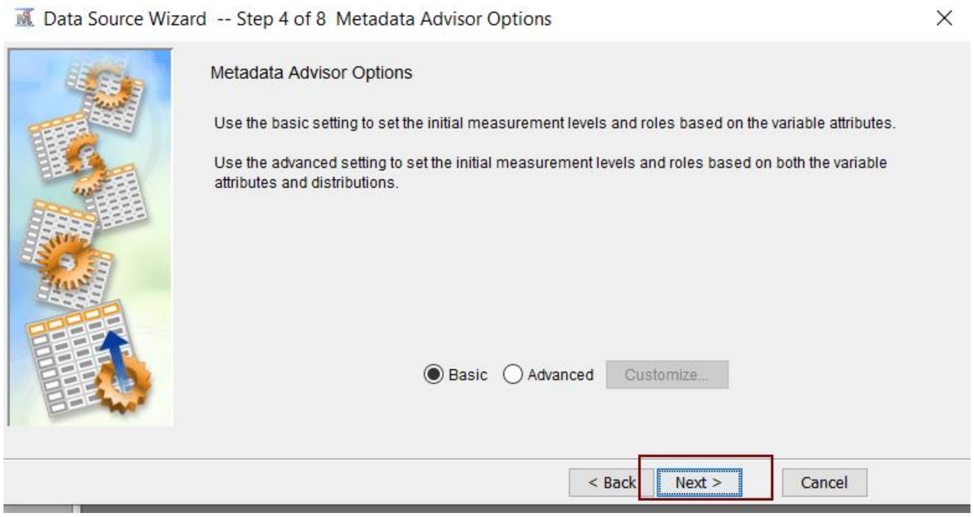
Table : DATA678.GYM_CROWD



Data Source Wizard -- Step 3 of 8 Table Information

Table Properties

Property	Value
Table Name	DATA678.GYM_CROWD
Description	
Member Type	DATA
Data Set Type	DATA
Engine	BASE
Number of Variables	12
Number of Observations	62184
Created Date	May 12, 2019 9:16:45 PM EDT
Modified Date	May 12, 2019 9:16:45 PM EDT



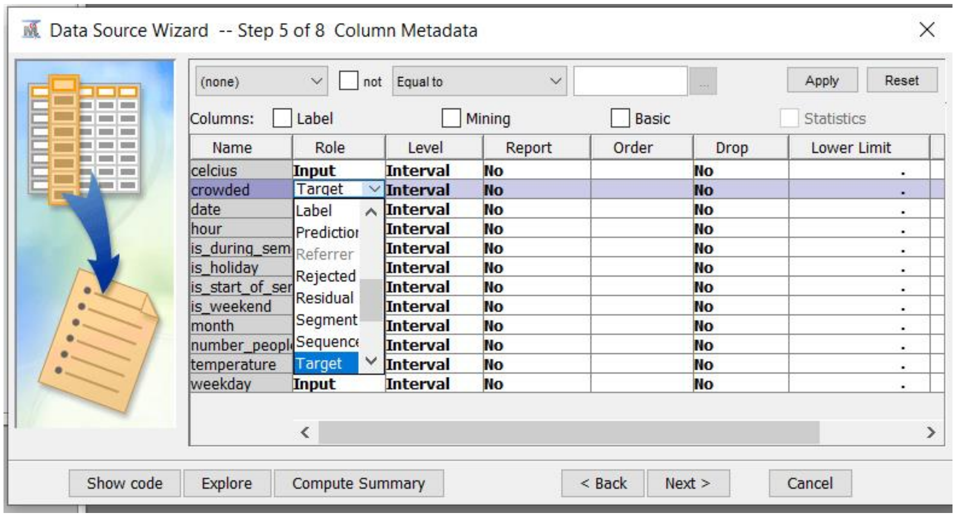
Data Source Wizard -- Step 4 of 8 Metadata Advisor Options

Metadata Advisor Options

Use the basic setting to set the initial measurement levels and roles based on the variable attributes.

Use the advanced setting to set the initial measurement levels and roles based on both the variable attributes and distributions.

● Basic ○ Advanced

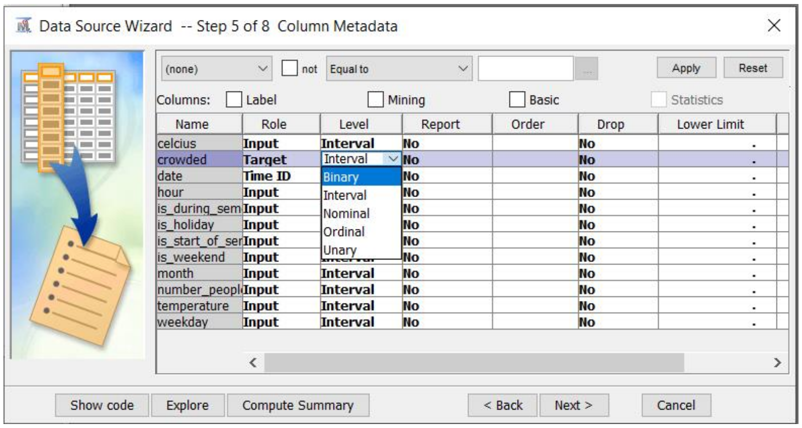


Data Source Wizard -- Step 5 of 8 Column Metadata

(none) □ not Equal to

Columns: □ Label □ Mining □ Basic □ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit
celcius	Input	Interval	No		No	.
crowded	Target	Interval	No		No	.
date		Interval	No		No	.
hour		Interval	No		No	.
is_during_sem		Interval	No		No	.
is_holiday		Interval	No		No	.
is_start_of_ser		Interval	No		No	.
is_weekend		Interval	No		No	.
month		Interval	No		No	.
number_peopl		Interval	No		No	.
temperature		Interval	No		No	.
weekday	Input	Interval	No		No	.

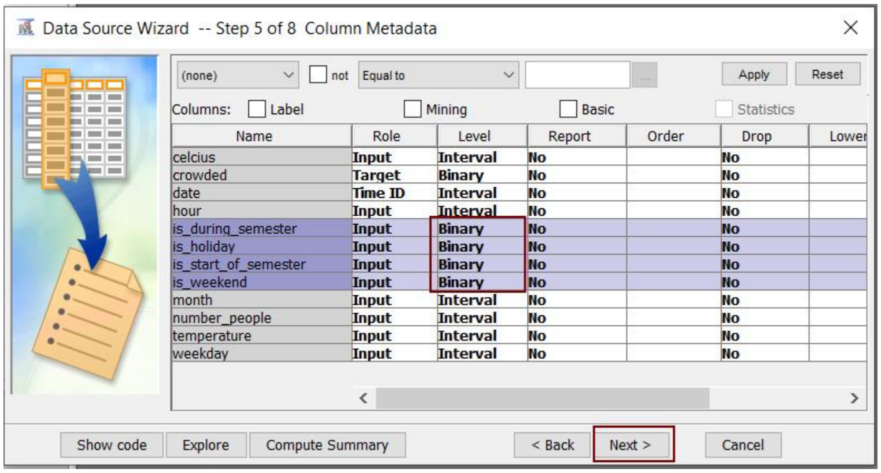


Data Source Wizard -- Step 5 of 8 Column Metadata

(none) □ not Equal to

Columns: □ Label □ Mining □ Basic □ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit
celcius	Input	Interval	No		No	.
crowded	Target	Interval	No		No	.
date	Time ID		No		No	.
hour	Input		No		No	.
is_during_sem	Input		No		No	.
is_holiday	Input		No		No	.
is_start_of_ser	Input		No		No	.
is_weekend	Input		No		No	.
month	Input	Interval	No		No	.
number_peopl	Input	Interval	No		No	.
temperature	Input	Interval	No		No	.
weekday	Input	Interval	No		No	.

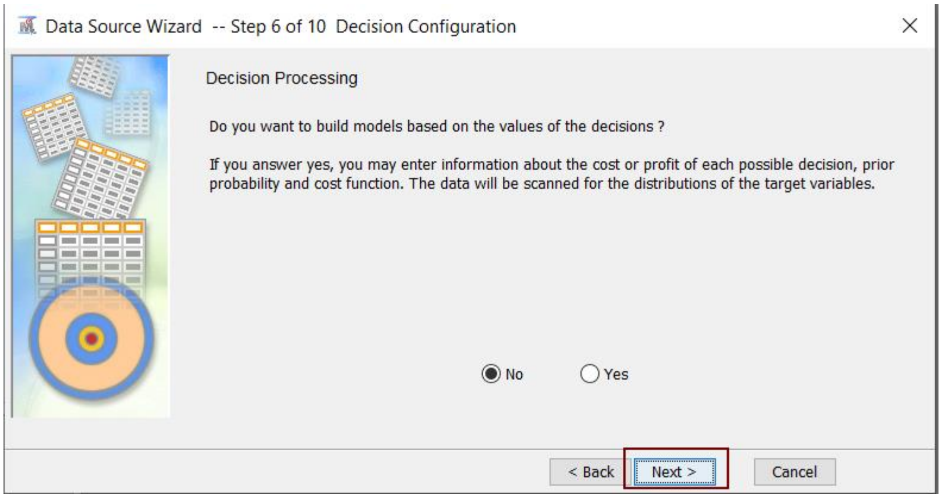


Data Source Wizard -- Step 5 of 8 Column Metadata

(none) □ not Equal to

Columns: □ Label □ Mining □ Basic □ Statistics

Name	Role	Level	Report	Order	Drop	Lower
celcius	Input	Interval	No		No	
crowded	Target	Binary	No		No	
date	Time ID	Interval	No		No	
hour	Input	Interval	No		No	
is_during_semester	Input	Binary	No		No	
is_holiday	Input	Binary	No		No	
is_start_of_semester	Input	Binary	No		No	
is_weekend	Input	Binary	No		No	
month	Input	Interval	No		No	
number_people	Input	Interval	No		No	
temperature	Input	Interval	No		No	
weekday	Input	Interval	No		No	



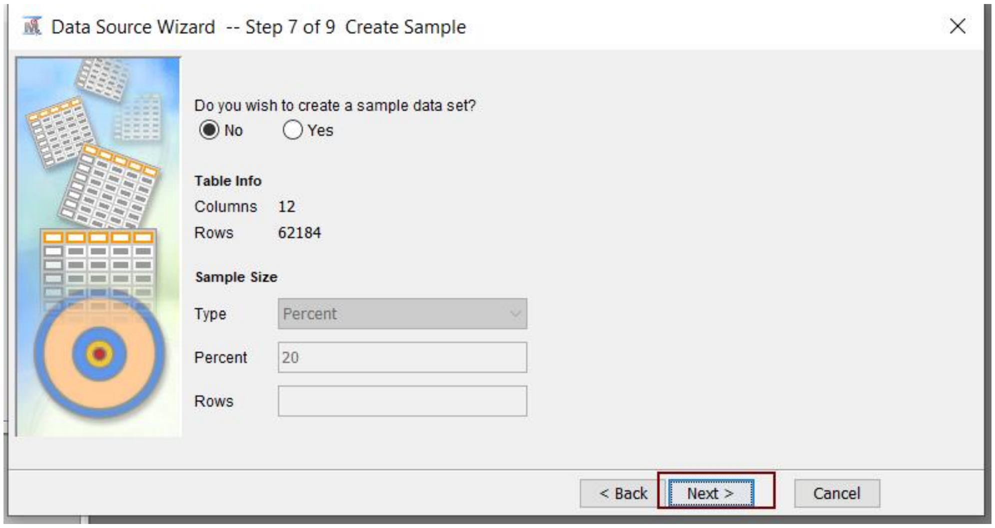
Data Source Wizard -- Step 6 of 10 Decision Configuration

Decision Processing

Do you want to build models based on the values of the decisions ?

If you answer yes, you may enter information about the cost or profit of each possible decision, prior probability and cost function. The data will be scanned for the distributions of the target variables.

● No ○ Yes



Data Source Wizard -- Step 7 of 9 Create Sample

Do you wish to create a sample data set?

● No ○ Yes

Table Info

Columns 12

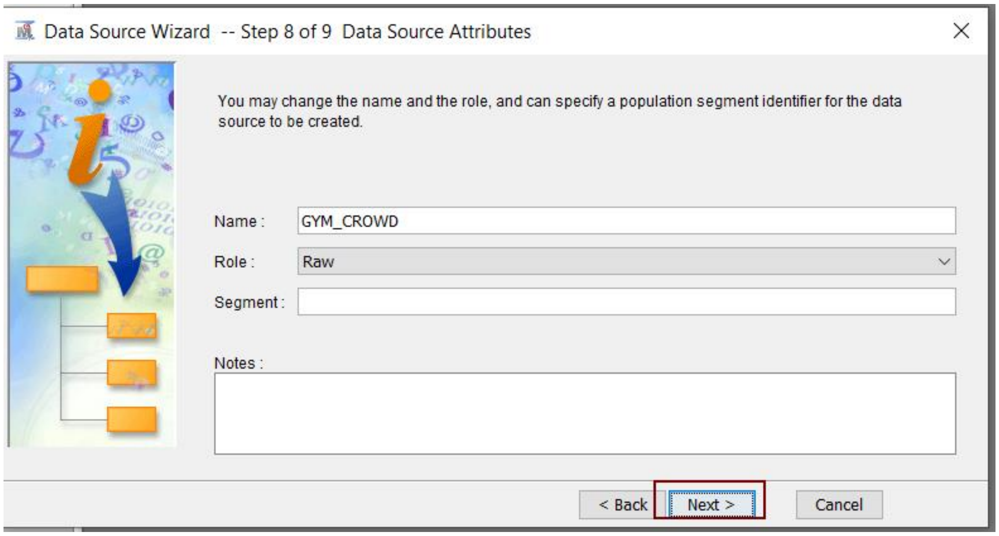
Rows 62184

Sample Size

Type Percent

Percent 20

Rows



Data Source Wizard -- Step 8 of 9 Data Source Attributes

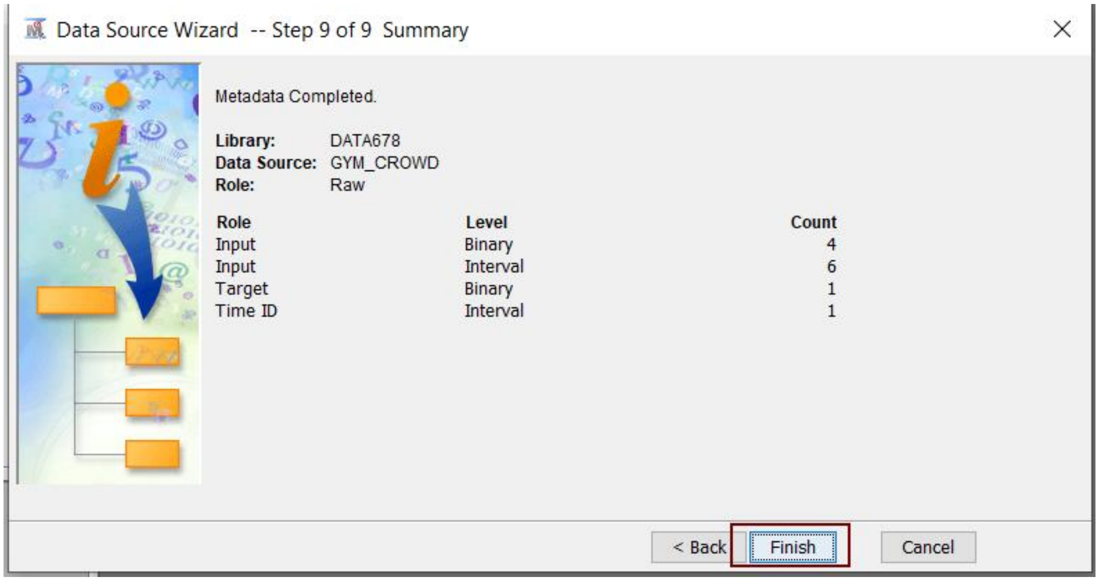
You may change the name and the role, and can specify a population segment identifier for the data source to be created.

Name : GYM_CROWD

Role : Raw

Segment :

Notes :



Data Source Wizard -- Step 9 of 9 Summary

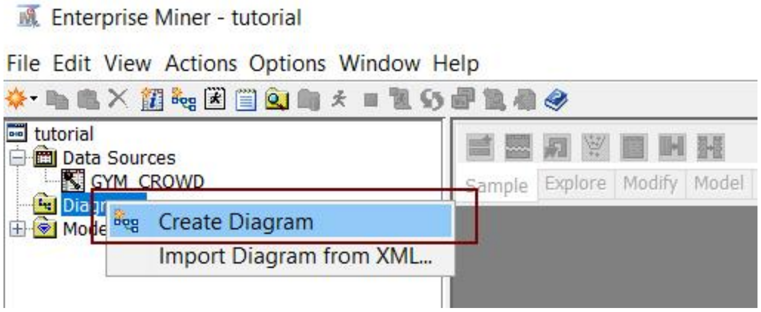
Metadata Completed.

Library: DATA678
Data Source: GYM_CROWD
Role: Raw

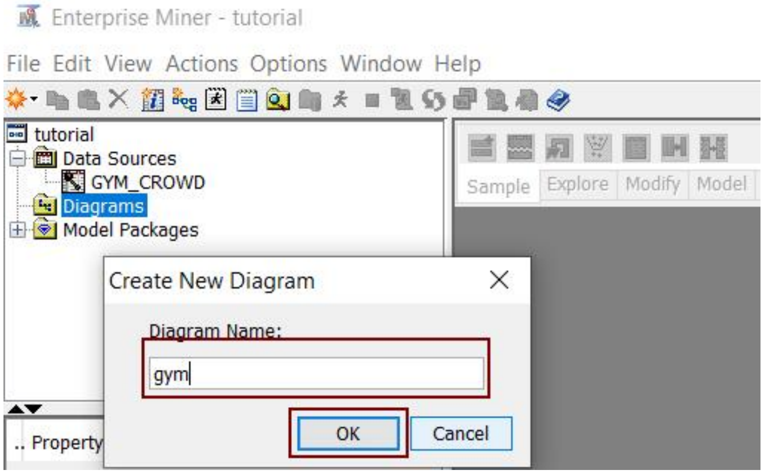
Role	Level	Count
Input	Binary	4
Input	Interval	6
Target	Binary	1
Time ID	Interval	1

4) CREATE A DIAGRAM:

right-click on DIAGRAM → <u>CREATE DIAGRAM → DIAGRAM NAME → OK</u>



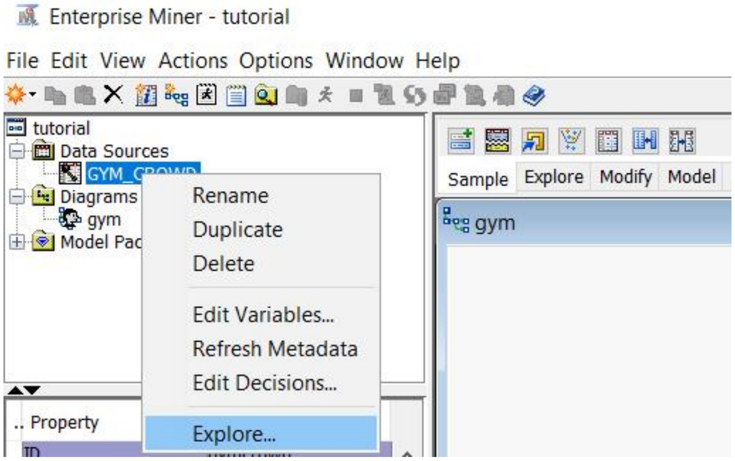
Enterprise Miner - tutorial
File Edit View Actions Options Window Help
tutorial
Data Sources
GYM_CROWD
Diagr
Mode
Create Diagram
Import Diagram from XML...
Sample
Explore
Modify
Model



HOW TO EXPLORE DATA

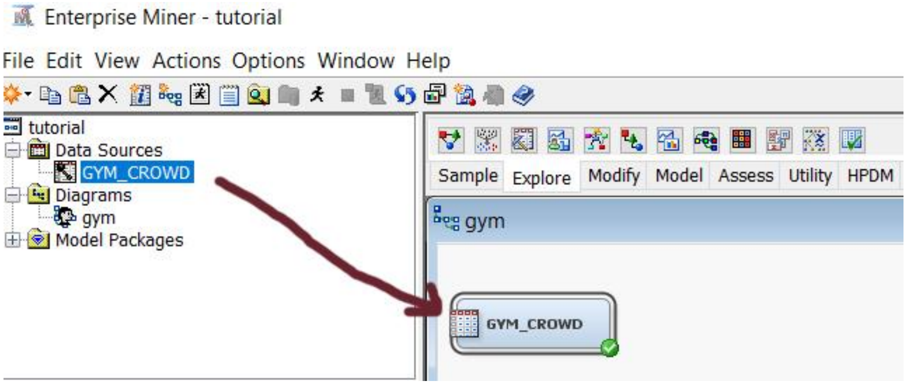
A) First way:

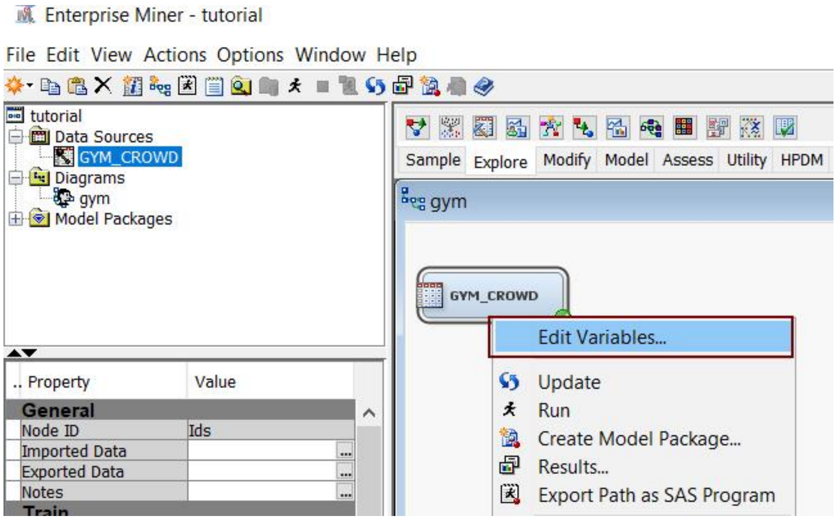
Right-click on the data source and choose EXPLORE



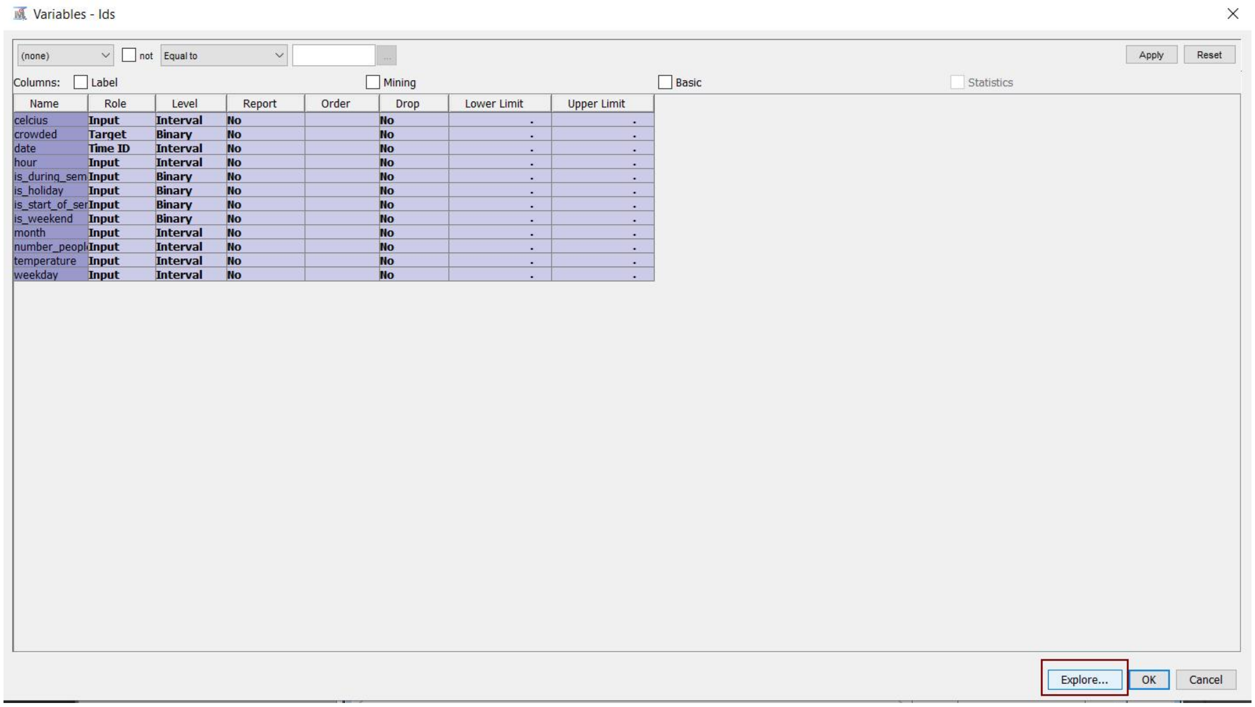
B) Second way:

OPEN THE DIAGRAM → DRAG THE DATA SOURCE INTO THE DIAGRAM → right-click on the file source you have just dragged → EDIT VARIABLES → SELECT THE VARIABLES → EXPLORE

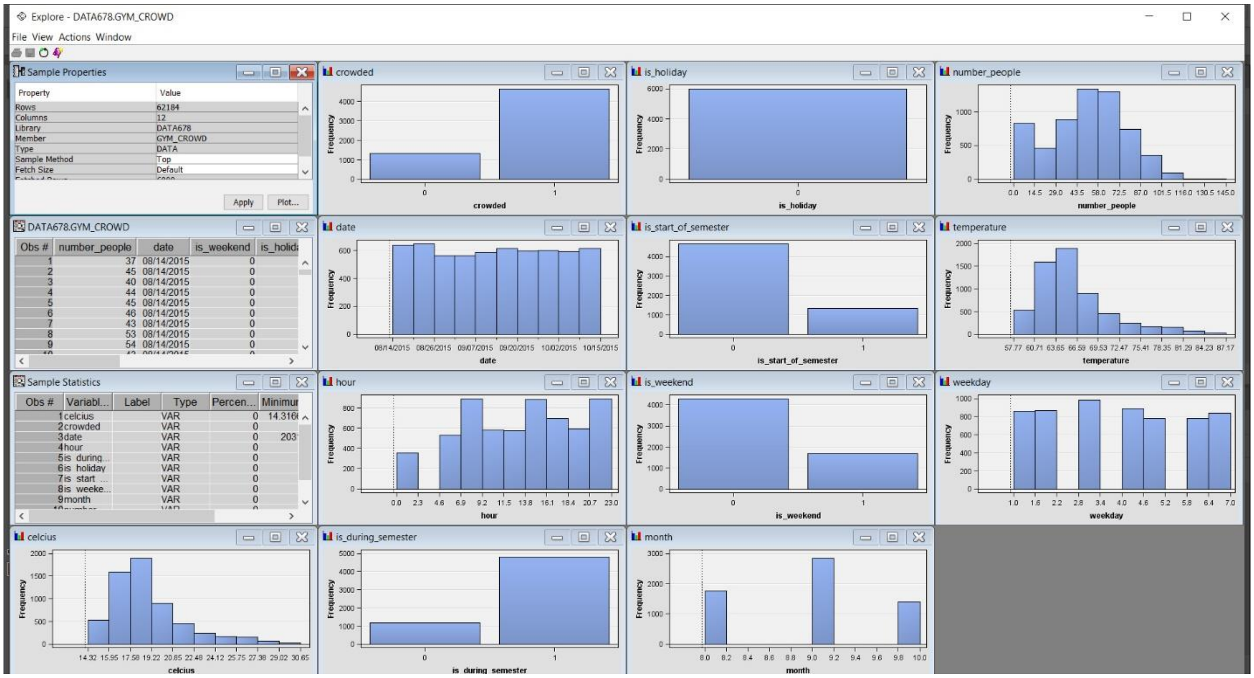




.. Property	Value
General	
Node ID	Ids
Imported Data	...
Exported Data	...
Notes	...
Train	



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
celcius	Input	Interval	No		No	.	.
crowded	Target	Binary	No		No	.	.
date	Time ID	Interval	No		No	.	.
hour	Input	Interval	No		No	.	.
is_during_sem	Input	Binary	No		No	.	.
is_holiday	Input	Binary	No		No	.	.
is_start_of_ser	Input	Binary	No		No	.	.
is_weekend	Input	Binary	No		No	.	.
month	Input	Interval	No		No	.	.
number_peopl	Input	Interval	No		No	.	.
temperature	Input	Interval	No		No	.	.
weekday	Input	Interval	No		No	.	.



Explore - DATA678.GYM_CROWD

Sample Properties

Property	Value
Rows	62184
Columns	12
Library	DATA678
Member	GYM_CROWD
Type	DATA
Sample Method	Top
Fetch Size	Default

DATA678.GYM_CROWD

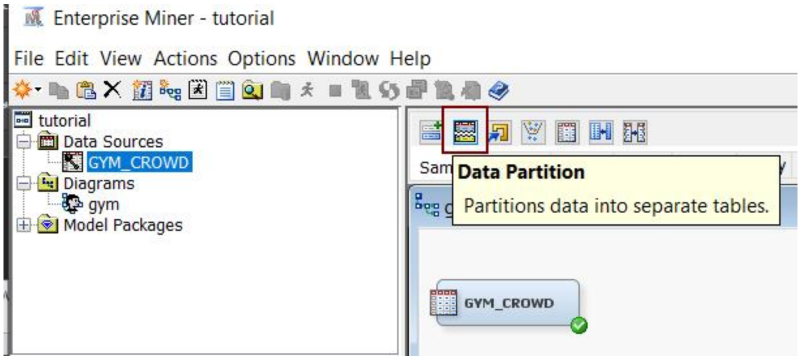
Obs #	number_people	date	is_weekend	is_holid
1	37	08/14/2015	0	
2	45	08/14/2015	0	
3	40	08/14/2015	0	
4	44	08/14/2015	0	
5	45	08/14/2015	0	
6	46	08/14/2015	0	
7	43	08/14/2015	0	
8	53	08/14/2015	0	
9	54	08/14/2015	0	

Sample Statistics

Obs #	Variabl...	Label	Type	Percen...	Minimur
1	celcius		VAR	0	14.316
2	crowded		VAR	0	
3	date		VAR	0	203
4	hour		VAR	0	
5	is_during...		VAR	0	
6	is_holiday		VAR	0	
7	is_start_...		VAR	0	
8	is_weeke...		VAR	0	
9	month		VAR	0	

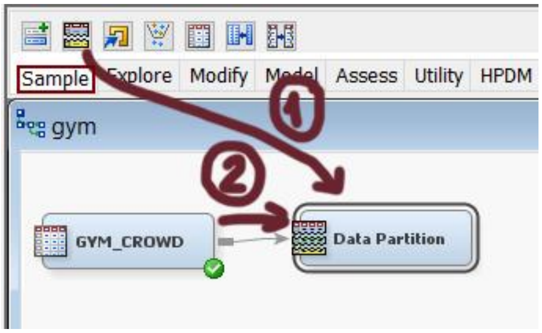
PARTIONING THE DATA: CREATING A TRAIN AND VALIDATION DATASET

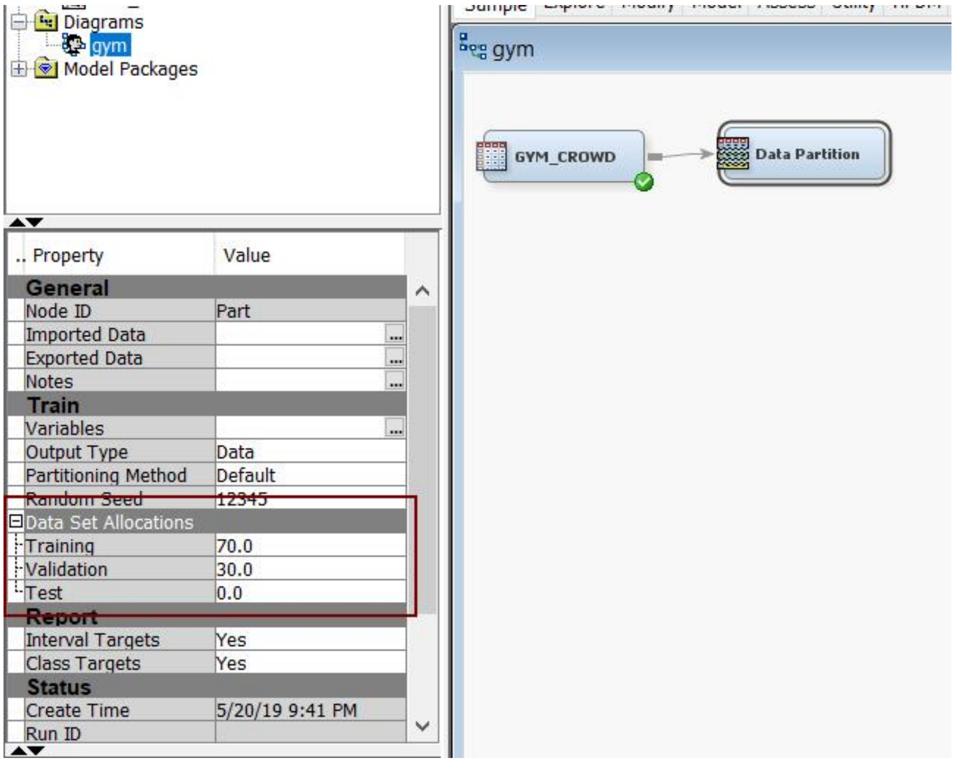
SAMPLE (ON THE RIGHT TAB) → DRAG AND DROP DATA PARTITION → CONNECT THE TABLE AND THE DATA PARTITION NODES → LOOK AT THE RIGHT OF THE DIAGRAM: DATA SET ALLOCATIONS: SET TRAINING TO 70% AND VALIDATION – 30% → right-click on the PARTITION NODE → RUN → YES → RESULTS (TO VIEW THE RESULTS OF THE PARTITION)



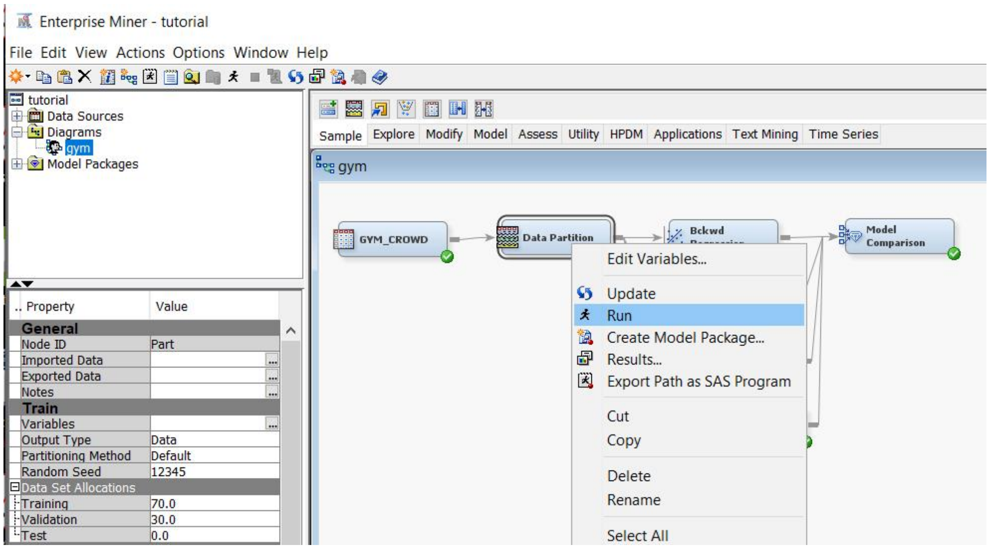
Data Partition
Partitions data into separate tables.

Enterprise Miner - tutorial
File Edit View Actions Options Window Help
tutorial
Data Sources
GYM_CROWD
Diagrams
gym
Model Packages
Sam
GYM_CROWD

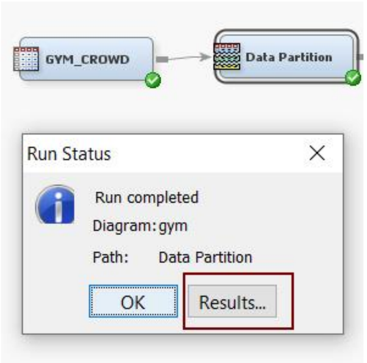




.. Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
⊟Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	5/20/19 9:41 PM
Run ID	



.. Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
⊟Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0



Run Status

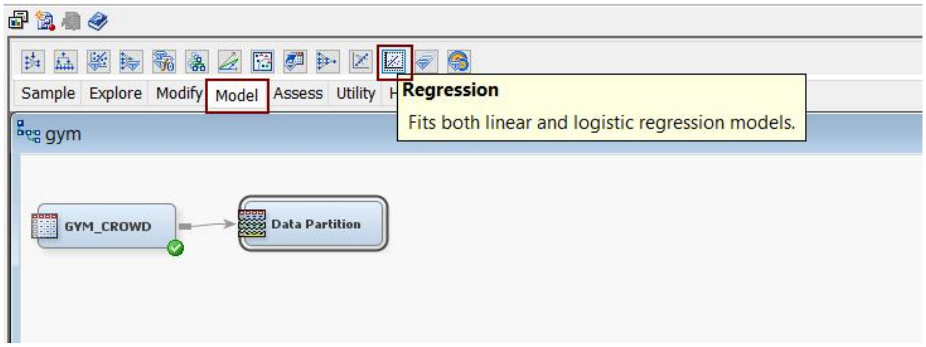
Run completed
Diagram: gym
Path: Data Partition

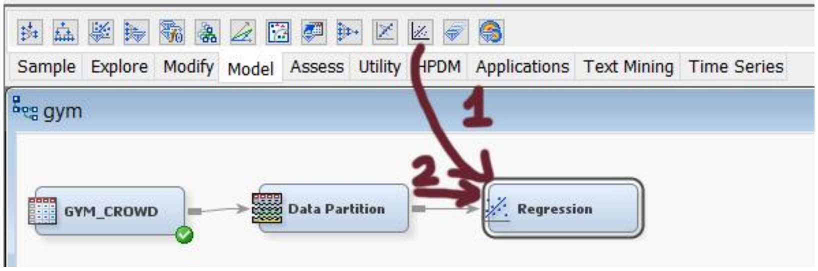
LOGISTIC REGRESSION

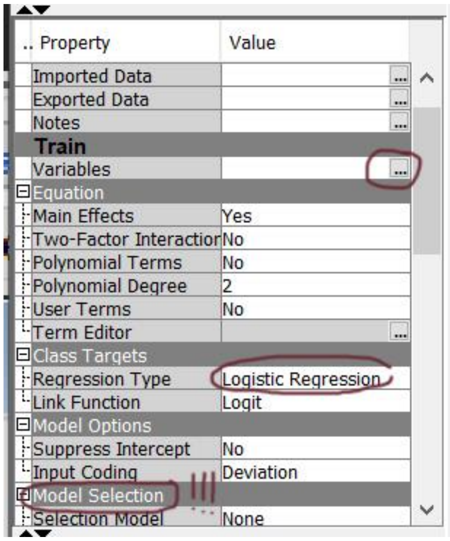
MODEL → REGRESSION → DRAG AND DROP THE REGRESSION TO THE DIAGRAM → CONNECT THE PARTITION TO THE REGRESSION NODE → YOU CAN EDIT VARIABLES IF YOU LIKE (ON THE LEFT; TRAIN) AND MAKE SURE THAT THE CLASS TRAGETS IS SET TO LOGISTIC REGRESSION → MODEL SELECTION: BACKWARD ETC. → Right-click on the regression node → RUN → YES → RESULTS

You will see 4 tables:

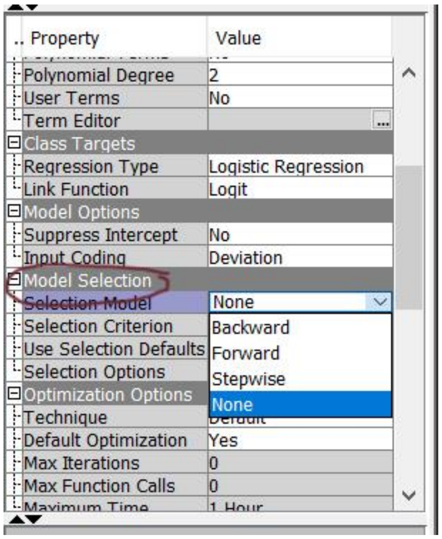
- **FIT STATISTICS will let you compare different values, i.e. MSE or some other errors for the TRAIN & VALIDATION data sets.**
- **OUTPUT: EVENT CLASSIFICATION TABLE (CONFUSION MATRIX)**
- **LIFT: CUMULATIVE AND REGULAR LIFT**
- **EFFECTS PLOT: TO EVALUATE WHICH VARIABLES HAVE THE HIGHEST IMPACT ON THE MODEL**



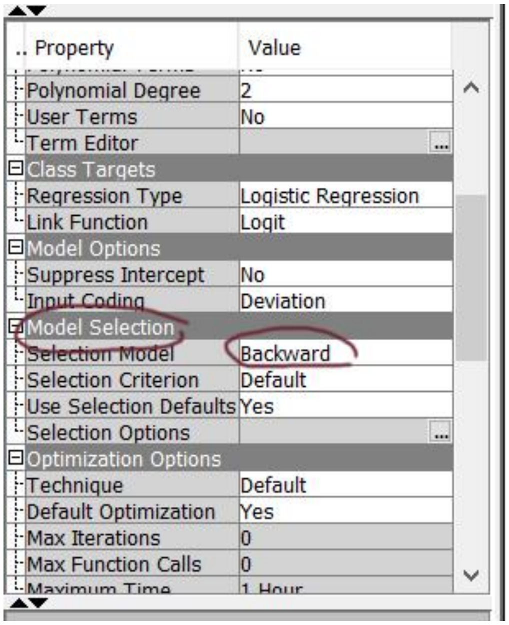




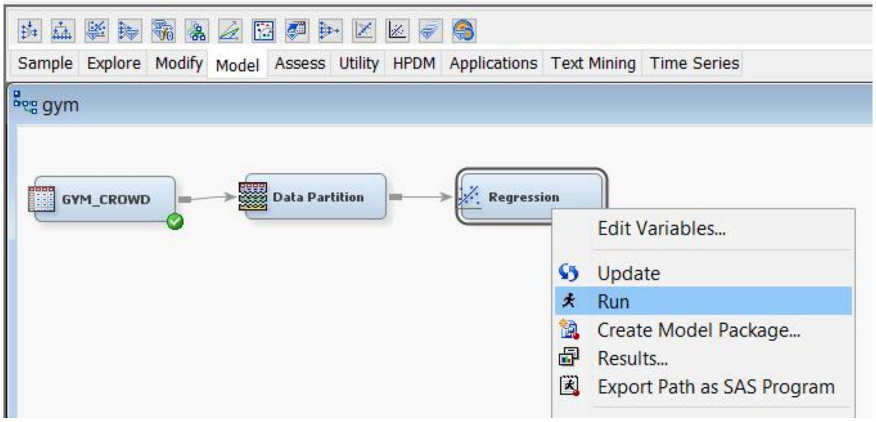
.. Property	Value
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactior	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None

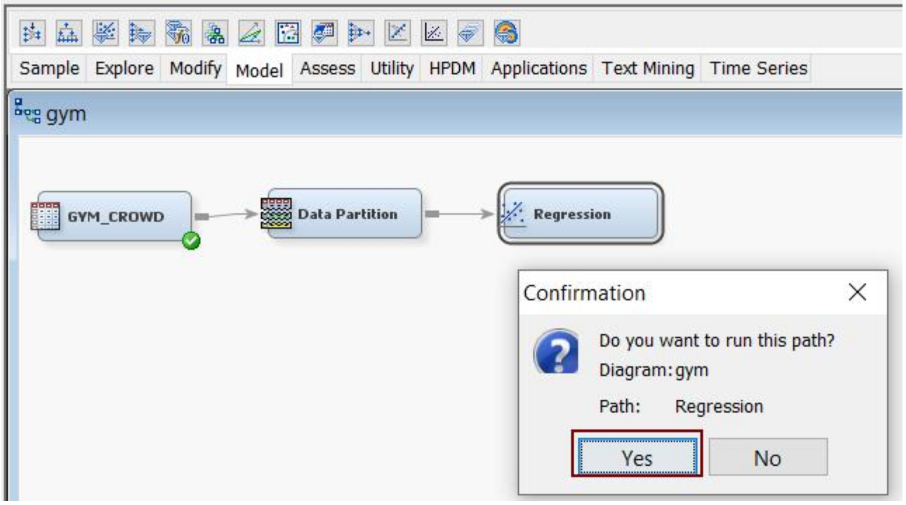


.. Property	Value
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Backward
Use Selection Defaults	Forward
Selection Options	Stepwise
Optimization Options	None
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour



.. Property	Value
Polynomial Degree	2
User Terms	No
Term Editor	...
⊟ Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
⊟ Model Options	
Suppress Intercept	No
Input Coding	Deviation
⊟ Model Selection	
Selection Model	Backward
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
⊟ Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour



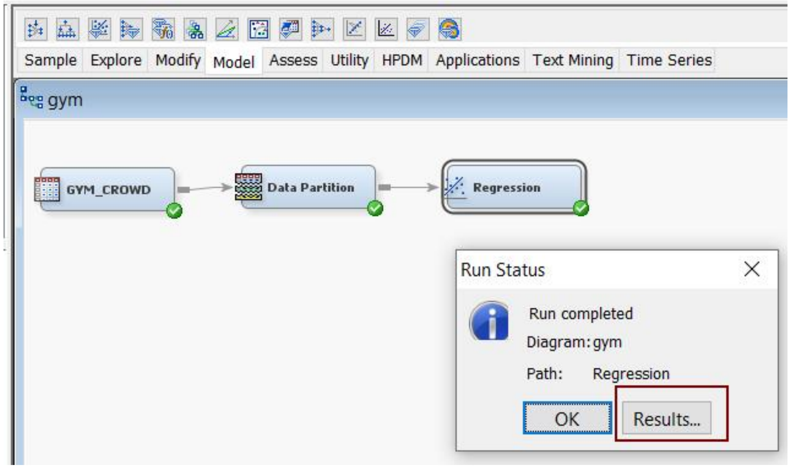


Confirmation

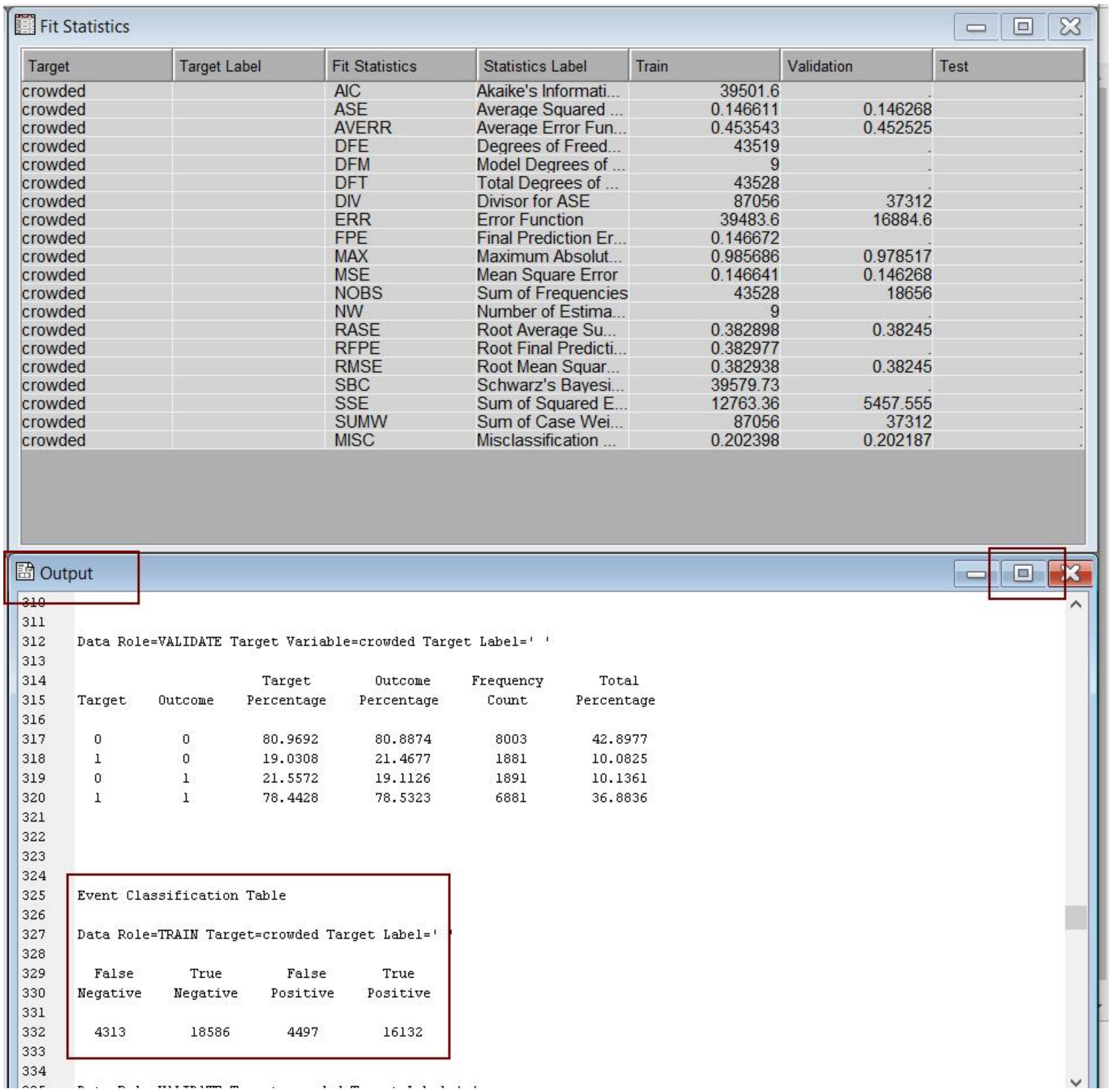
Do you want to run this path?

Diagram: gym

Path: Regression

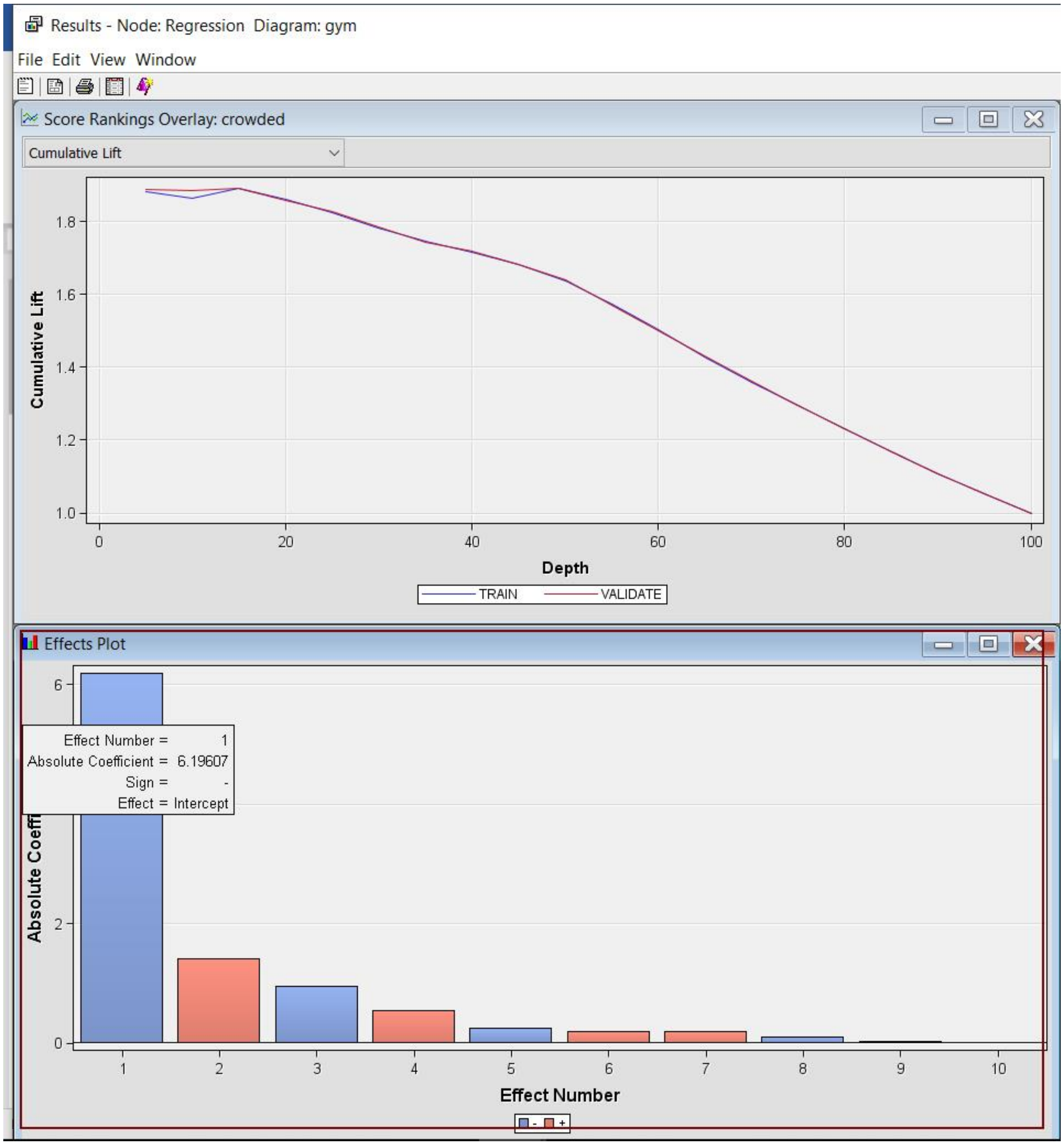


Run completed
Diagram: gym
Path: Regression



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
crowded		AIC	Akaike's Informati...	39501.6	.	.
crowded		ASE	Average Squared ...	0.146611	0.146268	.
crowded		AVERR	Average Error Fun...	0.453543	0.452525	.
crowded		DFE	Degrees of Freed...	43519	.	.
crowded		DFM	Model Degrees of ...	9	.	.
crowded		DFT	Total Degrees of ...	43528	.	.
crowded		DIV	Divisor for ASE	87056	37312	.
crowded		ERR	Error Function	39483.6	16884.6	.
crowded		FPE	Final Prediction Er...	0.146672	.	.
crowded		MAX	Maximum Absolut...	0.985686	0.978517	.
crowded		MSE	Mean Square Error	0.146641	0.146268	.
crowded		NOBS	Sum of Frequencies	43528	18656	.
crowded		NW	Number of Estima...	9	.	.
crowded		RASE	Root Average Su...	0.382898	0.38245	.
crowded		RFPE	Root Final Predicti...	0.382977	.	.
crowded		RMSE	Root Mean Squar...	0.382938	0.38245	.
crowded		SBC	Schwarz's Bayesi...	39579.73	.	.
crowded		SSE	Sum of Squared E...	12763.36	5457.555	.
crowded		SUMW	Sum of Case Wei...	87056	37312	.
crowded		MISC	Misclassification ...	0.202398	0.202187	.

```
Data Role=VALIDATE Target Variable=crowded Target Label=' '

                        Target       Outcome     Frequency      Total
Target    Outcome     Percentage    Percentage     Count      Percentage

  0          0         80.9692       80.8874       8003        42.8977
  1          0         19.0308       21.4677       1881        10.0825
  0          1         21.5572       19.1126       1891        10.1361
  1          1         78.4428       78.5323       6881        36.8836




Event Classification Table

Data Role=TRAIN Target=crowded Target Label=' '

  False       True       False       True
Negative    Negative    Positive    Positive

  4313       18586       4497       16132
```

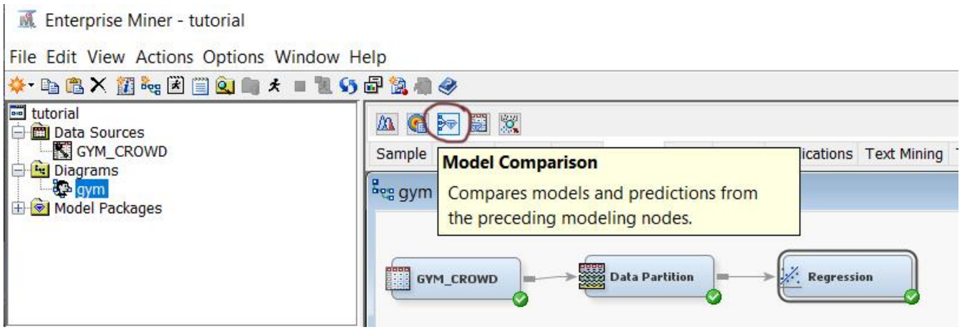


Effect Number = 1
Absolute Coefficient = 6.19607
Sign = -
Effect = Intercept

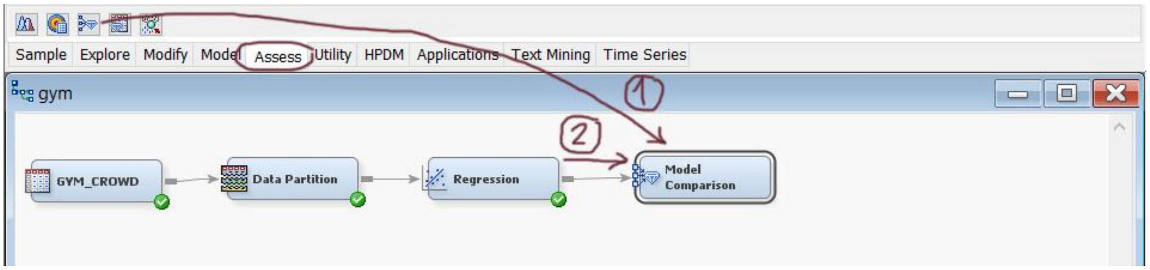
ASSESS A MODEL

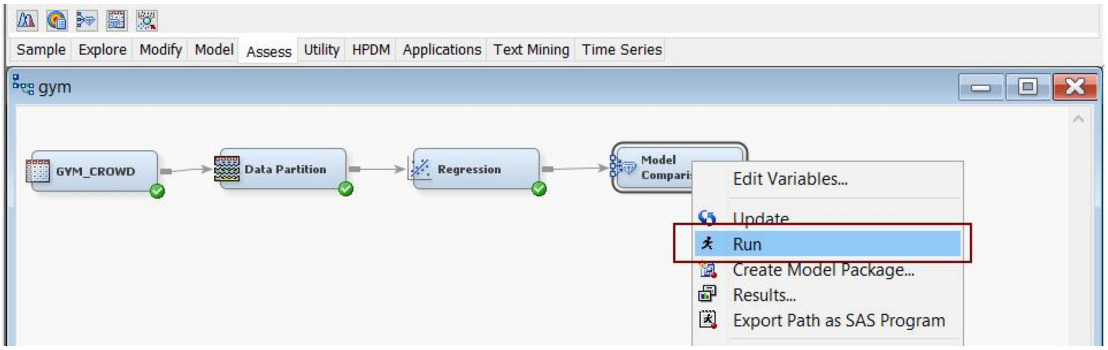
ASSESS → DRAG AND DROP: MODEL COMPARISON → CONNECT IT TO THE REGRESSION NODE → RUN → YES → RESULTS

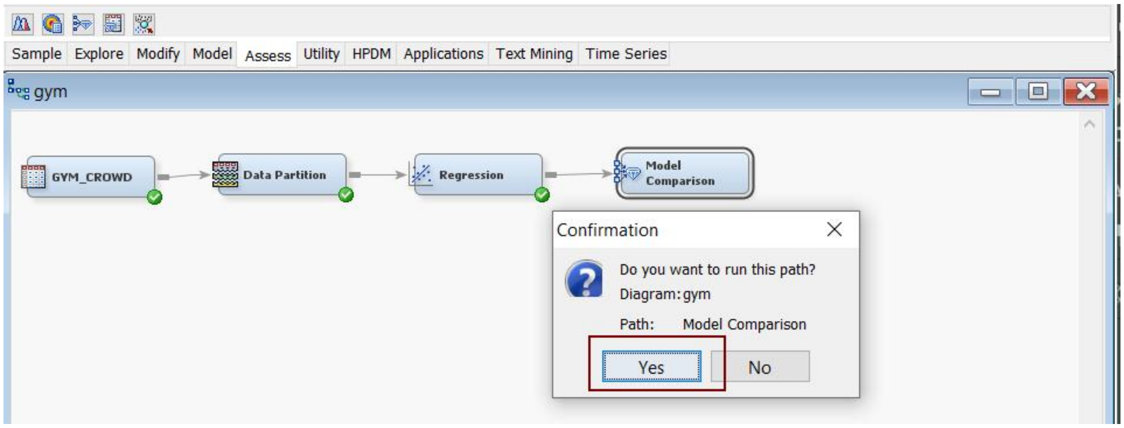
- **OUTPUT: Fit Statistics – it will provide you with the fit statistics of each model and if you compare a few models, then it will choose the best model based on the stats; you can also find other useful statistics**
- **LIFT: for train and validation data sets and if you have a few models you will see compared lifts for different models**
- **ROC CURVES**



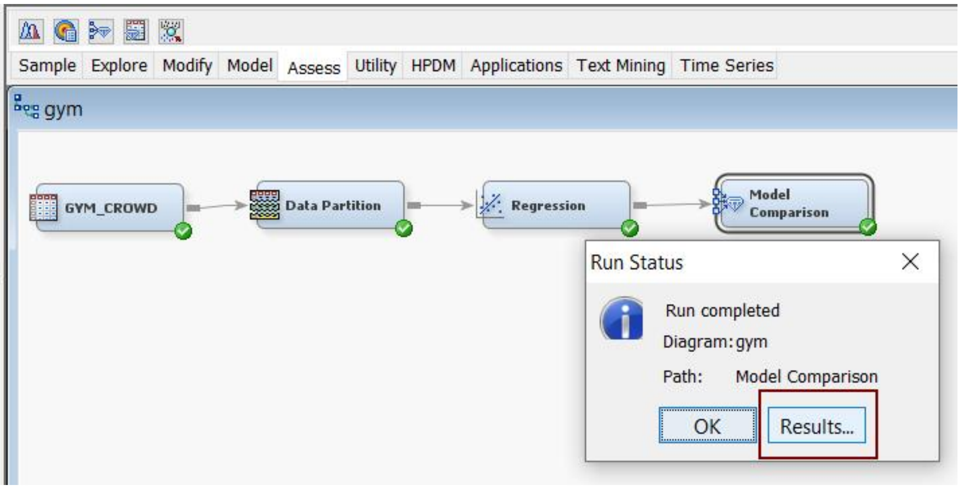
Model Comparison
Compares models and predictions from the preceding modeling nodes.





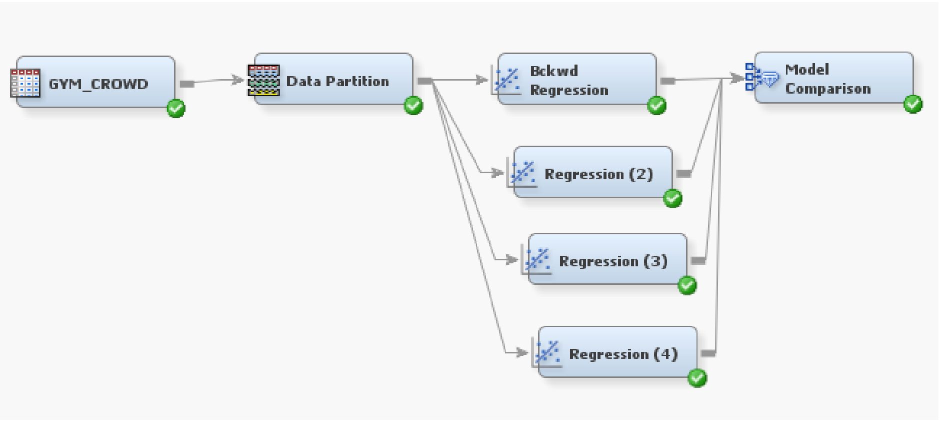


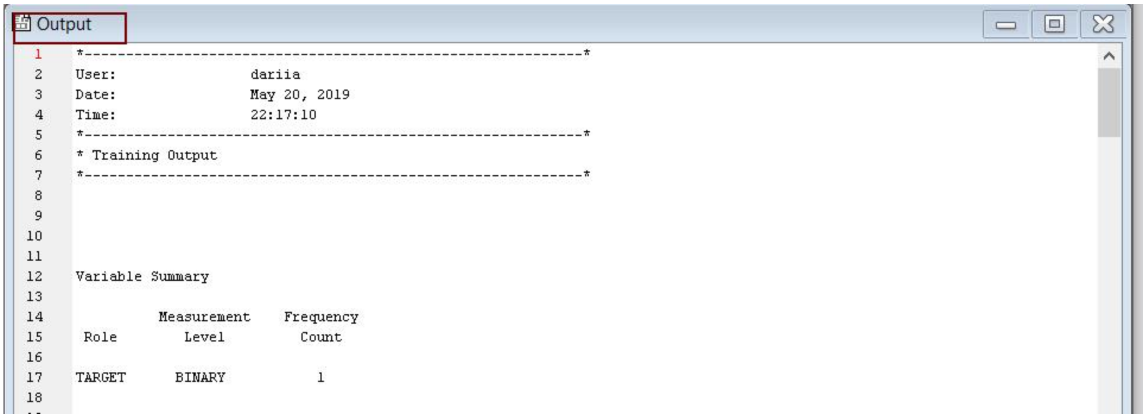
Do you want to run this path?
Diagram: gym
Path: Model Comparison



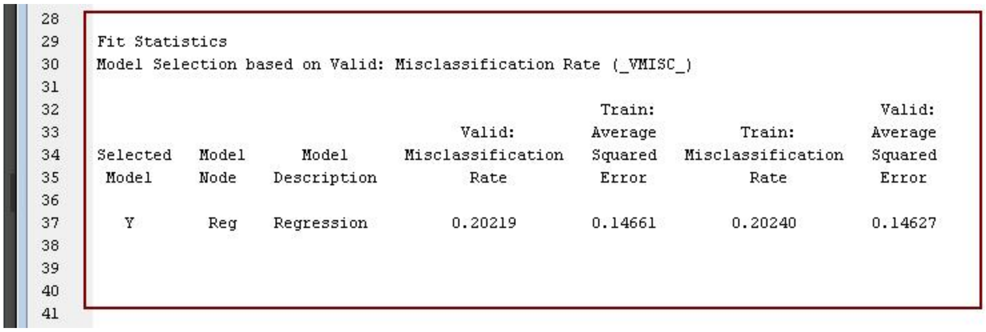
Run Status

Run completed
Diagram: gym
Path: Model Comparison

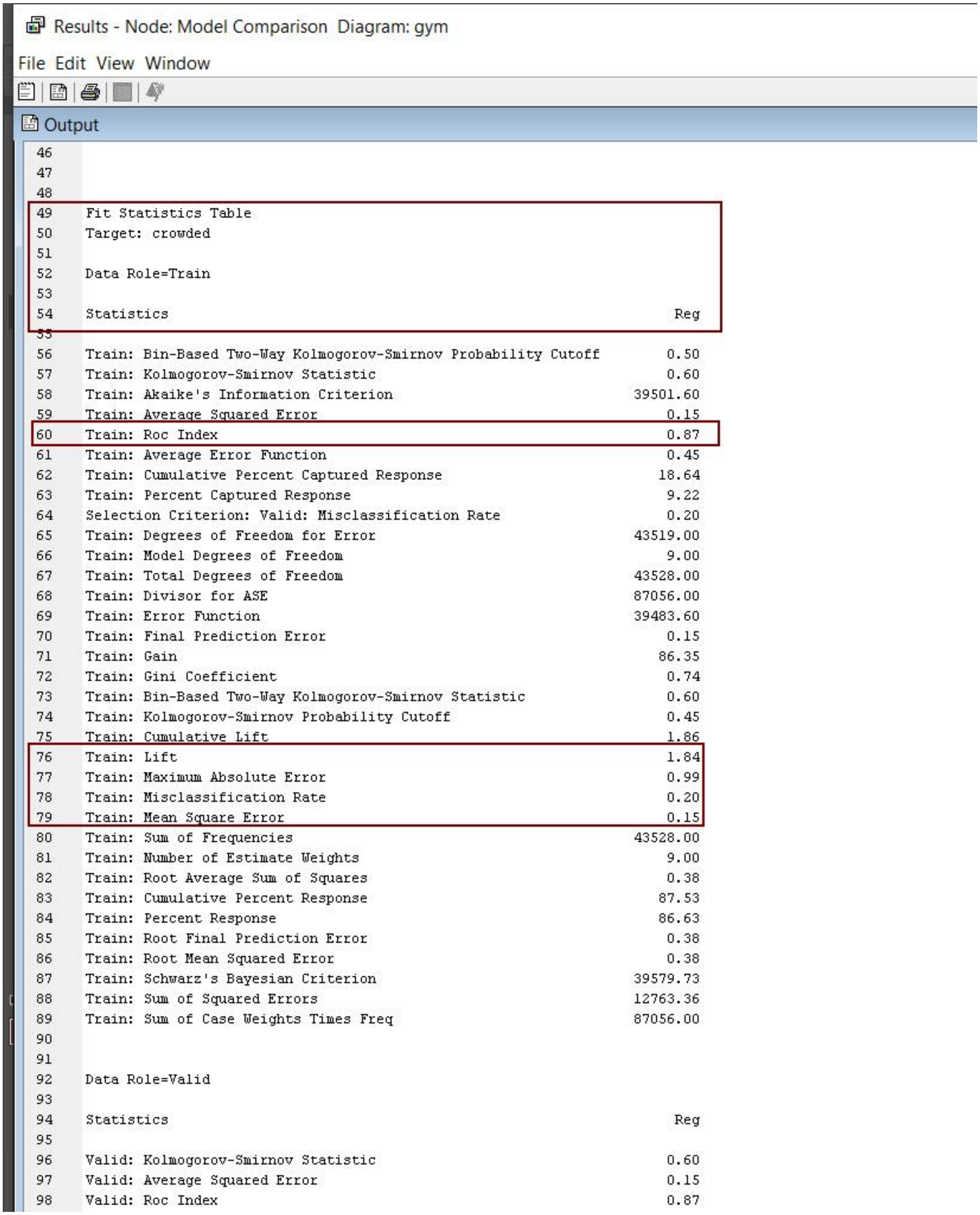




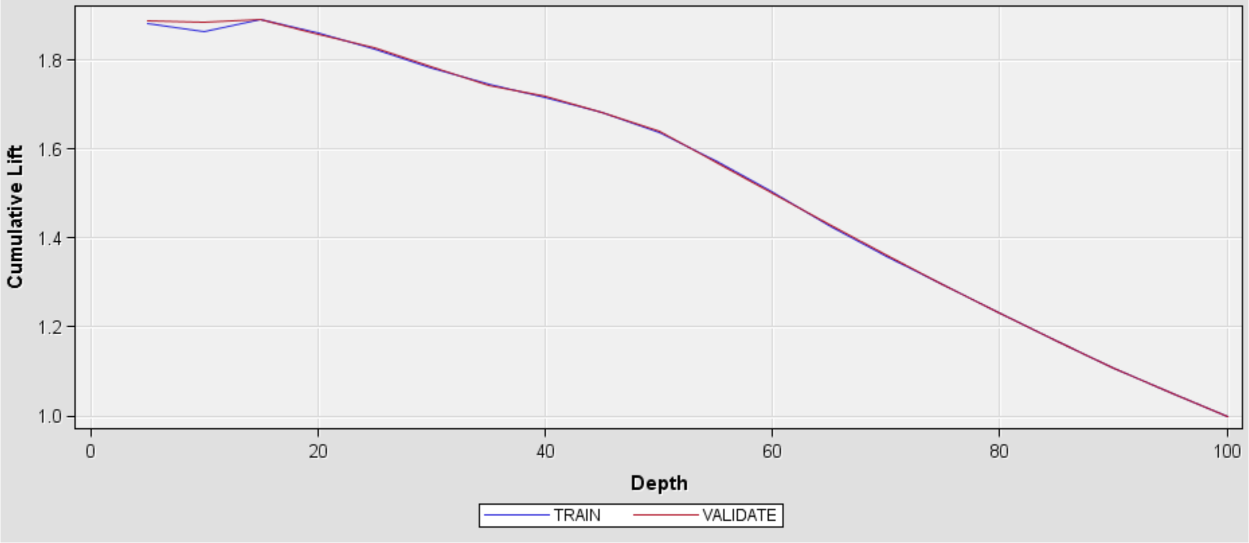
```
*------------------------------------------------------------*
User:                dariia
Date:                May 20, 2019
Time:                22:17:10
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

TARGET      BINARY           1

```

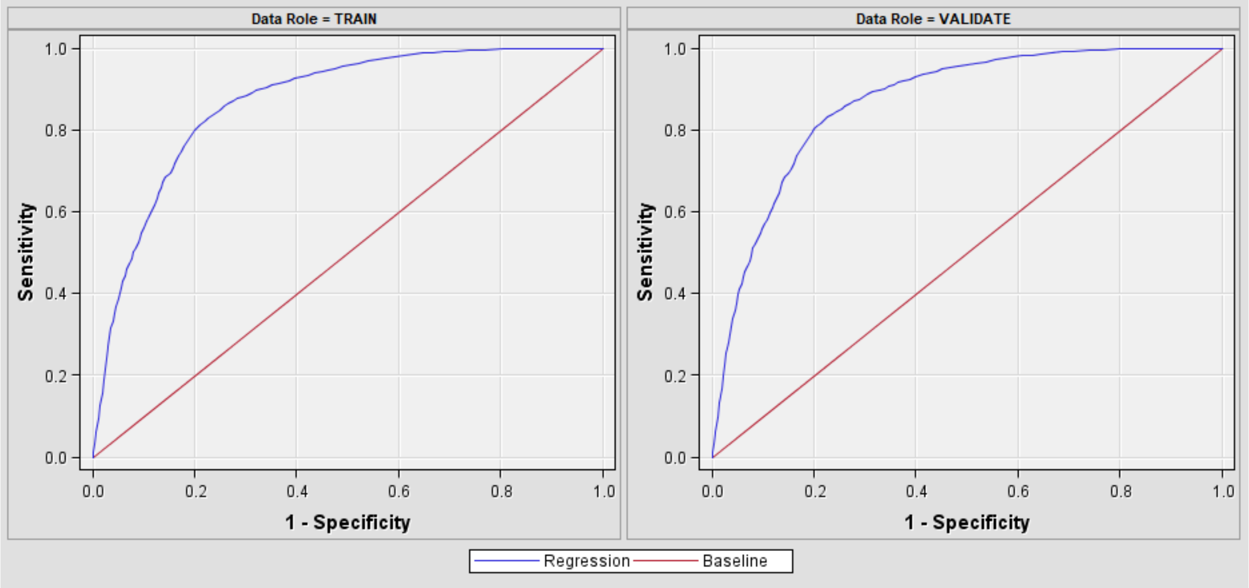


```
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

                                                   Train:                       Valid:
                                  Valid:           Average         Train:       Average
Selected   Model     Model     Misclassification   Squared   Misclassification  Squared
 Model     Node   Description        Rate           Error          Rate          Error

   Y        Reg    Regression       0.20219        0.14661        0.20240       0.14627



```



```
Fit Statistics Table
Target: crowded

Data Role=Train

Statistics                                                        Reg

Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff       0.50
Train: Kolmogorov-Smirnov Statistic                                  0.60
Train: Akaike's Information Criterion                            39501.60
Train: Average Squared Error                                         0.15
Train: Roc Index                                                     0.87
Train: Average Error Function                                        0.45
Train: Cumulative Percent Captured Response                         18.64
Train: Percent Captured Response                                     9.22
Selection Criterion: Valid: Misclassification Rate                   0.20
Train: Degrees of Freedom for Error                              43519.00
Train: Model Degrees of Freedom                                      9.00
Train: Total Degrees of Freedom                                  43528.00
Train: Divisor for ASE                                           87056.00
Train: Error Function                                            39483.60
Train: Final Prediction Error                                        0.15
Train: Gain                                                         86.35
Train: Gini Coefficient                                              0.74
Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic                0.60
Train: Kolmogorov-Smirnov Probability Cutoff                         0.45
Train: Cumulative Lift                                               1.86
Train: Lift                                                          1.84
Train: Maximum Absolute Error                                        0.99
Train: Misclassification Rate                                        0.20
Train: Mean Square Error                                             0.15
Train: Sum of Frequencies                                        43528.00
Train: Number of Estimate Weights                                    9.00
Train: Root Average Sum of Squares                                   0.38
Train: Cumulative Percent Response                                  87.53
Train: Percent Response                                             86.63
Train: Root Final Prediction Error                                   0.38
Train: Root Mean Squared Error                                       0.38
Train: Schwarz's Bayesian Criterion                              39579.73
Train: Sum of Squared Errors                                     12763.36
Train: Sum of Case Weights Times Freq                            87056.00


Data Role=Valid

Statistics                                                        Reg

Valid: Kolmogorov-Smirnov Statistic                                  0.60
Valid: Average Squared Error                                         0.15
Valid: Roc Index                                                     0.87
```


Cumulative Lift
1.8
1.6
1.4
1.2
1.0
0
20
40
60
80
100
Depth
TRAIN
VALIDATE


Data Role = TRAIN
Data Role = VALIDATE
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
Regression
Baseline

Questions:

Logistic regression

1. Create a new project, assign a library, create a data source and a new diagram.
2. Open the file called gym_crowd
3. Assign variable roles and levels
4. Explore the data in two ways and answer the questions:
    a. What distribution does the number of visitors have?
    b. What is the minimum and maximum number of gym visitors?
    c. How many variables and how many observations does the dataset have?
    d. Are there any variables that have missing observations? What should you do with missing values?
    e. Which variables do you think should be excluded from the analysis since they will not contribute much to the explanation of the number of visitors to the gym?
    f. Are there any redundant variables?
    g. Which variable should you exclude from your analysis?
5. Partition the data into TRAIN (70%) and VALIDATION (30%) datasets
6. Run a full logistic regression for event = 1:
    a. Which variables are the most important in the model?
    b. What is the model's cumulative lift? Interpret it.
    c. Use a Model Comparison node to get a ROC curve. What is AUC? Interpret it.
7. Run the following logistic regressions for event=1:
    a. Stepwise
    b. Forward
    c. Backward
8. Compare all four models. Which model is the best? Which criteria have you used?
9. Using the output found after running a stepwise regression model, create a **CONFUSION MATRIX** in **EXCEL** and calculate the following:
    a. True positive rate (TPR)
    b. False negative rate (FNR)
    c. True negative rate (TNR)
    d. False positive rate (FPR)
    e. Misclassification rate
    f. Sensitivity (TRUE POSITIVE RATE)
    g. Specificity (TRUE NEGATIVE RATE)
    h. Horizontal ROC coordinates
    i. Vertical ROC coordinates
    j. Event rate
    k. Actual rate among predicted
    l. Lift
    m. Gain