Overview of k-Means

Dr. Goutam Chakraborty

Outline

- Differences between hierarchical and nonhierarchical clustering methods.
- Advantages and disadvantages of both types of methods.
- Mechanics of k-Means clustering

Hierarchical versus Nonhierarchical Clustering Methods

Hierarchical

- Involves a *tree-like* construction process where clusters at any level of the tree are a combination of clusters below that level.
- After an observation has joined another observation in a step, successive steps keep them together.

Nonhierarchical

- Assigns objects into *prespecified* number of clusters using a distance/similarity metric. No tree-like structure exists.
- Assignment of object to cluster is *not fixed* through the iteration process.
- Iterate to minimize or maximize a criterion such as separation between clusters or, within-cluster similarity.

Advantages and Disadvantages of Hierarchical Methods

- Advantages include:
  - **Ability to capture non-spherical clusters**.
  - No order effect, that is, the ordering of observations has no impact on cluster solutions.
  - No need to make an initial guess at number of clusters in the data.
- Disadvantages include:
  - **Does not scale well for large/complex data**.
  - Early combinations (even if it is a mistake) persist throughout the process.
  - Susceptible to outliers (depends on method).
  - In many segmentation studies, there is little theoretical reason to expect a hierarchical structure.
  - Too many choices of methods.

Advantages and Disadvantages of Nonhierarchical Methods

- Advantages include:
  - **Scale up well with large/complex data**.
  - Generally easy to understand.
- Disadvantages include:
  - **Makes assumptions about shape of clusters**.
  - Number of clusters need to be specified in advance.
  - Results might be influenced by the choice of initial seeds or, order of reading of seeds.
  - Susceptible to outliers.

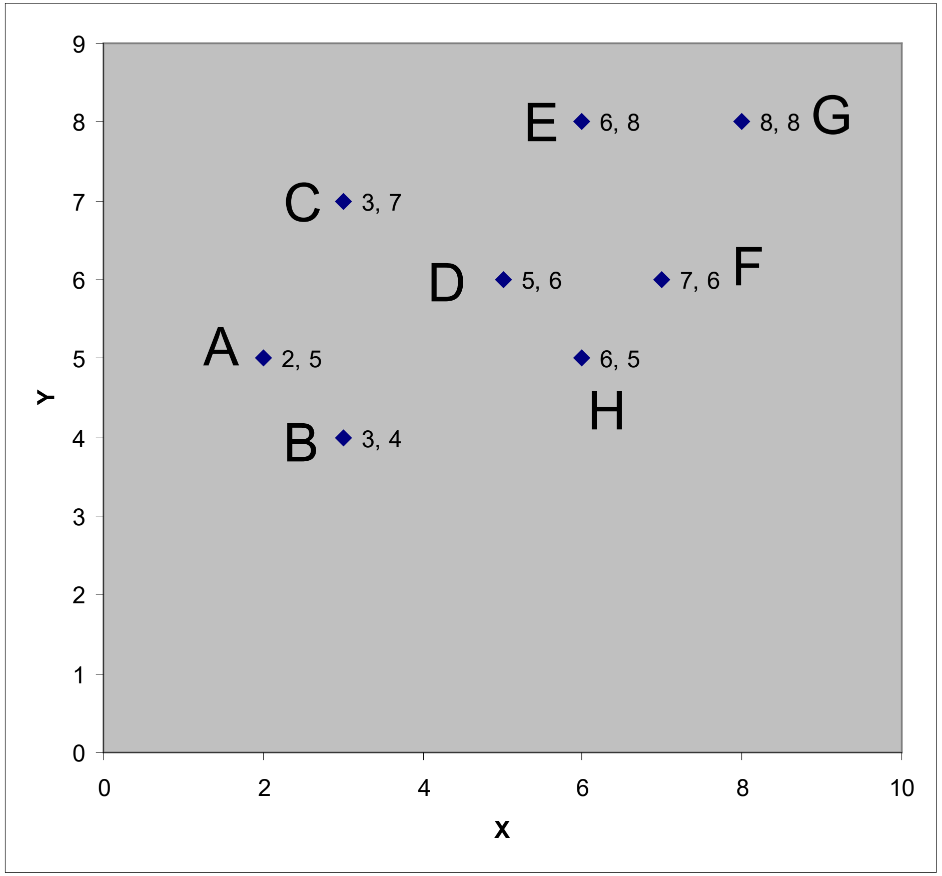
k-Means Procedure (Mechanics)

1. Select *k* cluster centers.
2. Assign cases to closest center.
3. Update cluster centers.
4. Reassign cases.
5. Repeat steps 3 and 4 until convergence.

A Numerical Example of *k*-Means Clustering

- Data from eight subjects (A, B, C, D, E, F, G, H) on two variables, X and Y.

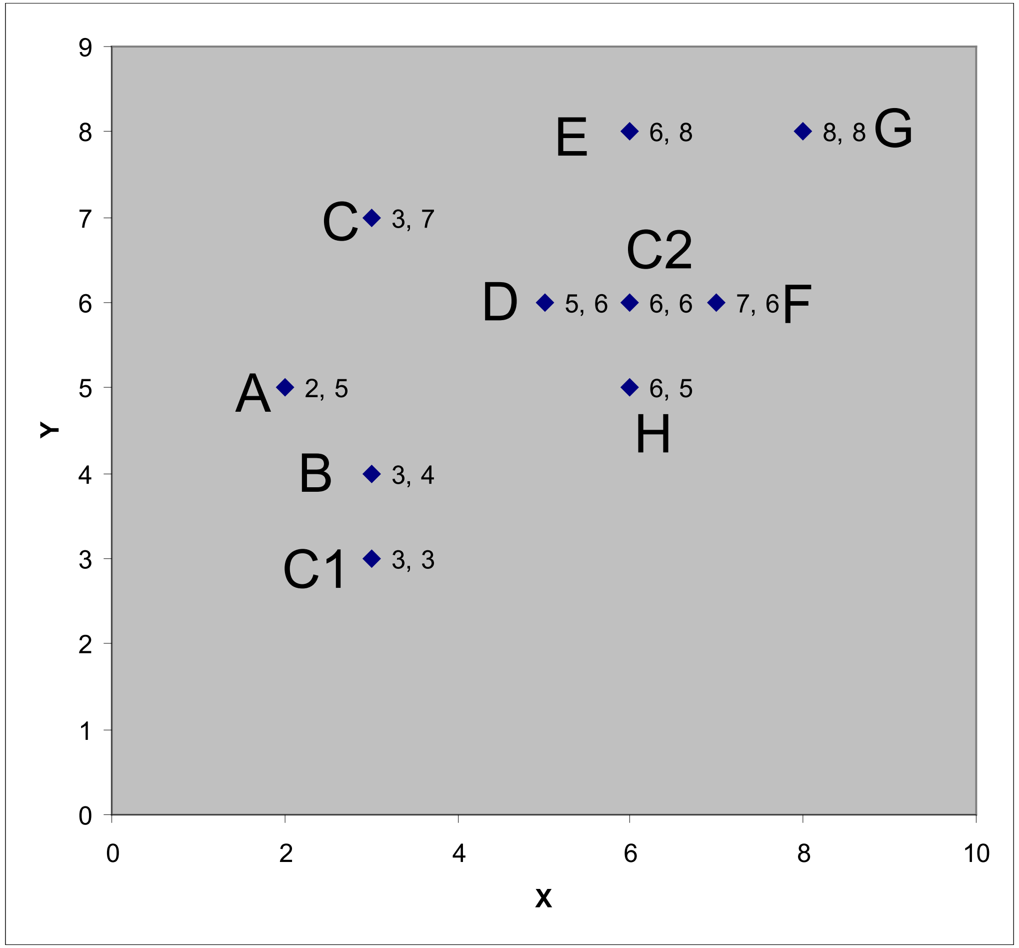
ID	X	Y
A	2	5
B	3	4
C	3	7
D	5	6
E	6	8
F	7	6
G	8	8
H	6	5



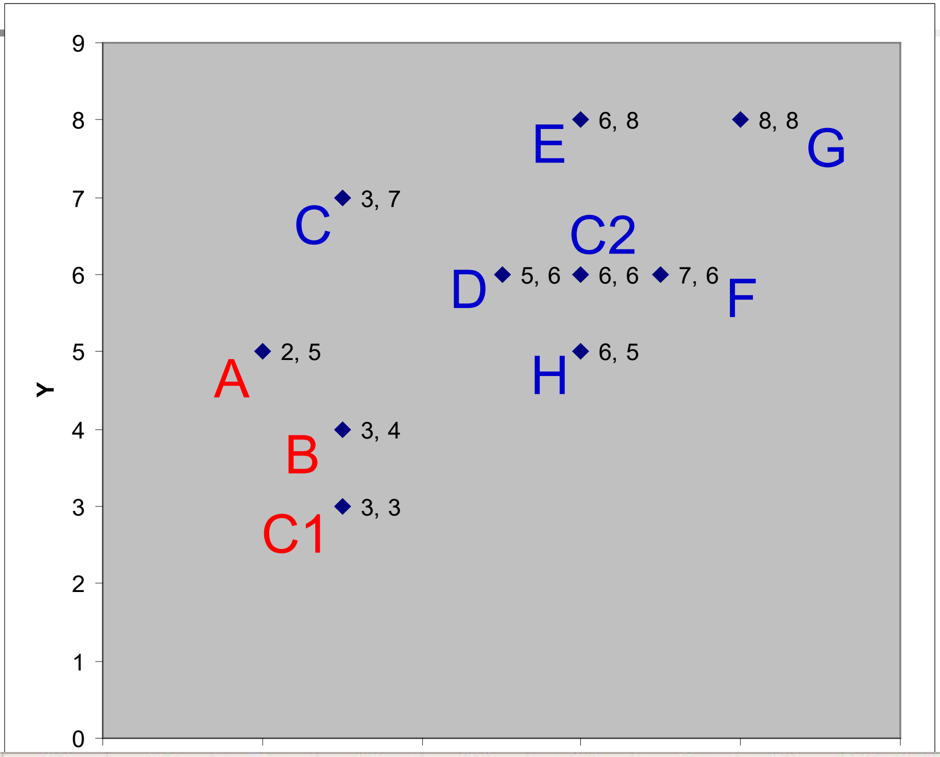
A Numerical Example of *k*-Means Clustering

- Center 1, C1(3,3) and Center 2, C2 (6,6) chosen at random.

ID	X	Y
A	2	5
B	3	4
C	3	7
D	5	6
E	6	8
F	7	6
G	8	8
H	6	5
Center1	3	3
Center2	6	6



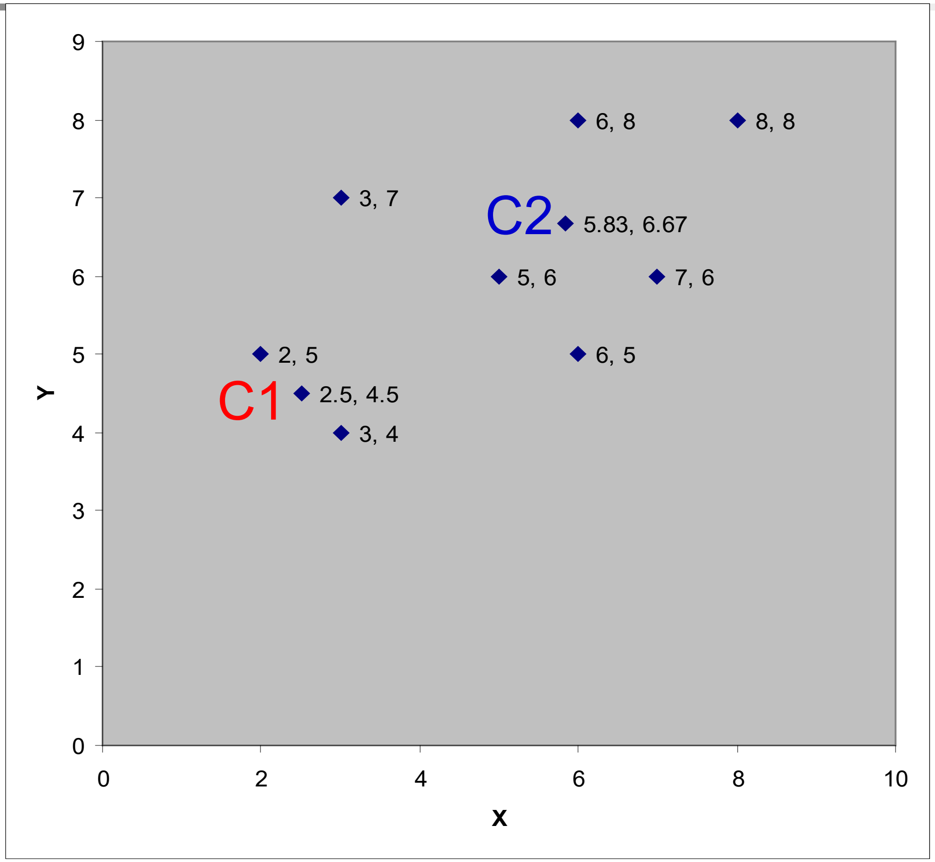
k-Means Numerical Example (Iteration 1)



ID	A	B	C	D	E	F	G	H	Center1	Center2
A	0.00000	.	.	.	.	.	.	.	.	.
B	1.41421	0.00000	.	.	.	.	.	.	.	.
C	2.23607	3.00000	0.00000	.	.	.	.	.	.	.
D	3.16228	2.82843	2.23607	0.00000	.	.	.	.	.	.
E	5.00000	5.00000	3.16228	2.23607	0.00000	.	.	.	.	.
F	5.09902	4.47214	4.12311	2.00000	2.23607	0.00000	.	.	.	.
G	6.70820	6.40312	5.09902	3.60555	2.00000	2.23607	0.00000	.	.	.
H	4.00000	3.16228	3.60555	1.41421	3.00000	1.41421	3.60555	0.00000	.	.
Center1	2.23607	1.00000	4.00000	3.60555	5.83095	5.00000	7.07107	3.60555	0.00000	.
Center2	4.12311	3.60555	3.16228	1.00000	2.00000	1.00000	2.82843	1.00000	4.24264	0

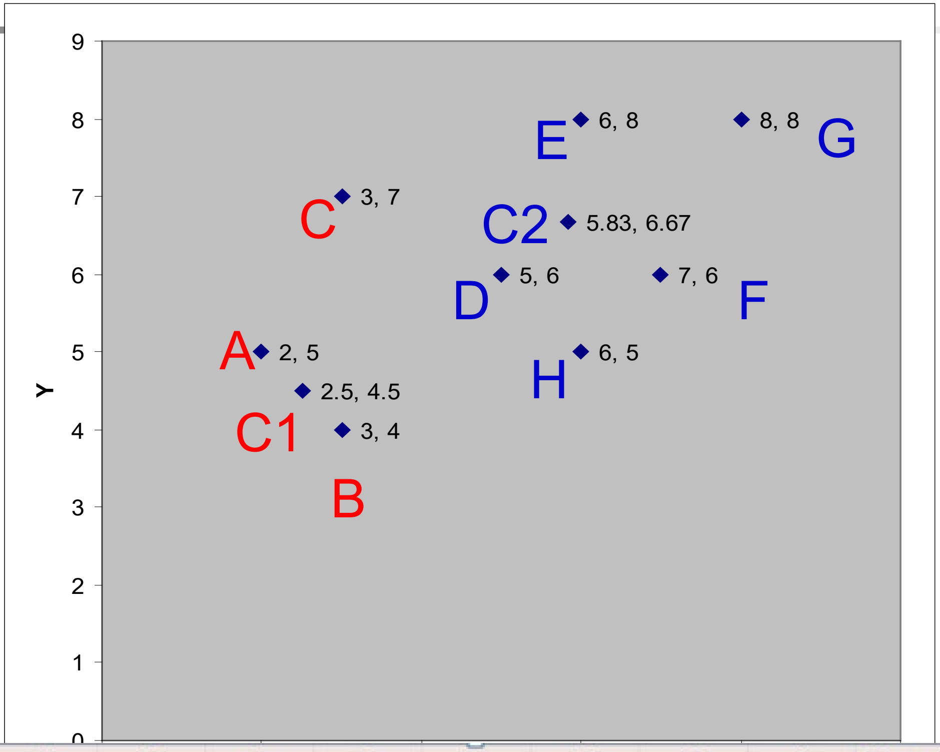
continued...

k-Means Numerical Example (Iteration 1)



- Update cluster centers (C1 and C2).
- New centers are C1(2.5,4.5) and C2(5.83,6.67).

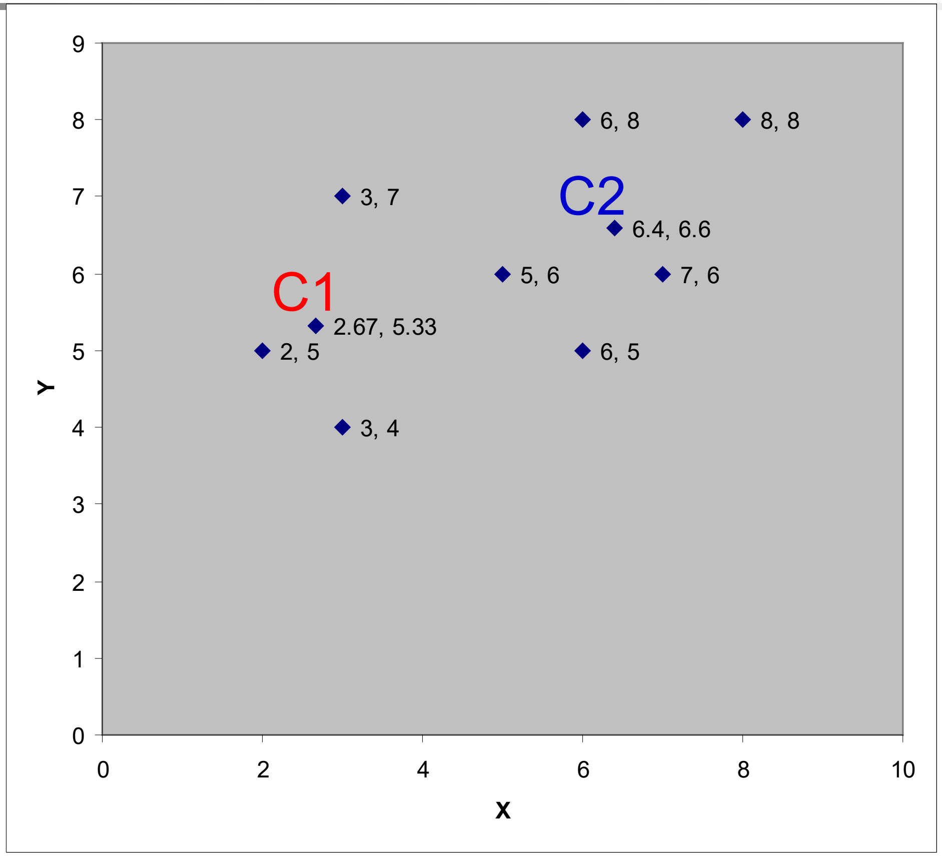
k-Means Numerical Example (Iteration 2)



ID	A	B	C	D	E	F	G	H	Center1	Center2
A	0.00000	.	.	.	.	.	.	.	.	.
B	1.41421	0.00000	.	.	.	.	.	.	.	.
C	2.23607	3.00000	0.00000	.	.	.	.	.	.	.
D	3.16228	2.82843	2.23607	0.00000	.	.	.	.	.	.
E	5.00000	5.00000	3.16228	2.23607	0.00000	.	.	.	.	.
F	5.09902	4.47214	4.12311	2.00000	2.23607	0.00000	.	.	.	.
G	6.70820	6.40312	5.09902	3.60555	2.00000	2.23607	0.00000	.	.	.
H	4.00000	3.16228	3.60555	1.41421	3.00000	1.41421	3.60555	0.00000	.	.
Center1	0.70711	0.70711	2.54951	2.91548	4.94975	4.74342	6.51920	3.53553	0.00000	.
Center2	4.17825	3.89073	2.84918	1.06668	1.34082	1.34826	2.54515	1.67863	3.97464	0

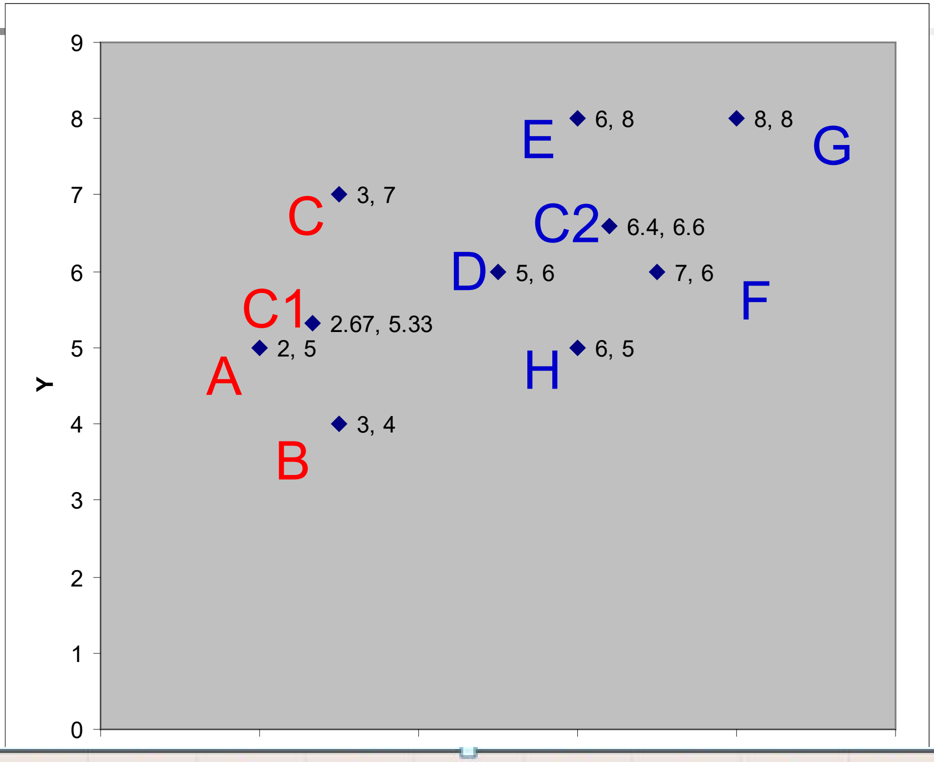
continued...

k-Means Numerical Example (Iteration 2)



- Update cluster centers (C1 and C2).
- New centers are C1(2.67,5.33) and C2(6.4,6.6)

k-Means Numerical Example (Iteration 3)



ID	A	B	C	D	E	F	G	H	Center1	Center2
A	0.00000	.	.	.	.	.	.	.	.	.
B	1.41421	0.00000	.	.	.	.	.	.	.	.
C	2.23607	3.00000	0.00000	.	.	.	.	.	.	.
D	3.16228	2.82843	2.23607	0.00000	.	.	.	.	.	.
E	5.00000	5.00000	3.16228	2.23607	0.00000	.	.	.	.	.
F	5.09902	4.47214	4.12311	2.00000	2.23607	0.00000	.	.	.	.
G	6.70820	6.40312	5.09902	3.60555	2.00000	2.23607	0.00000	.	.	.
H	4.00000	3.16228	3.60555	1.41421	3.00000	1.41421	3.60555	0.00000	.	.
Center1	0.74686	1.37033	1.70229	2.42442	4.26823	4.38153	5.96136	3.34631	0.00000	.
Center2	4.68188	4.28019	3.42345	1.52315	1.45602	0.84853	2.12603	1.64924	3.94028	0

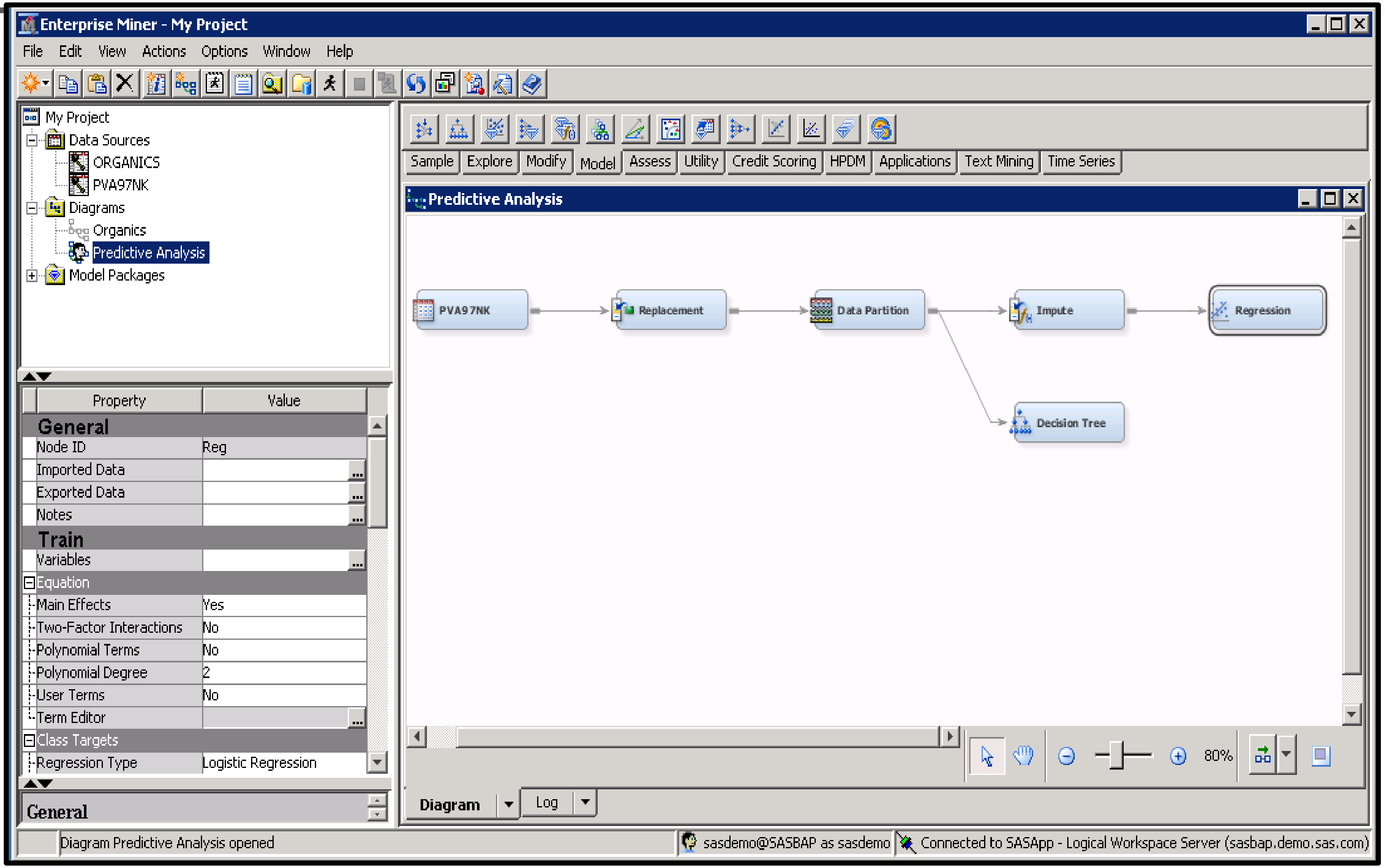
SAS EM Interface Tour

Dr. Goutam Chakraborty

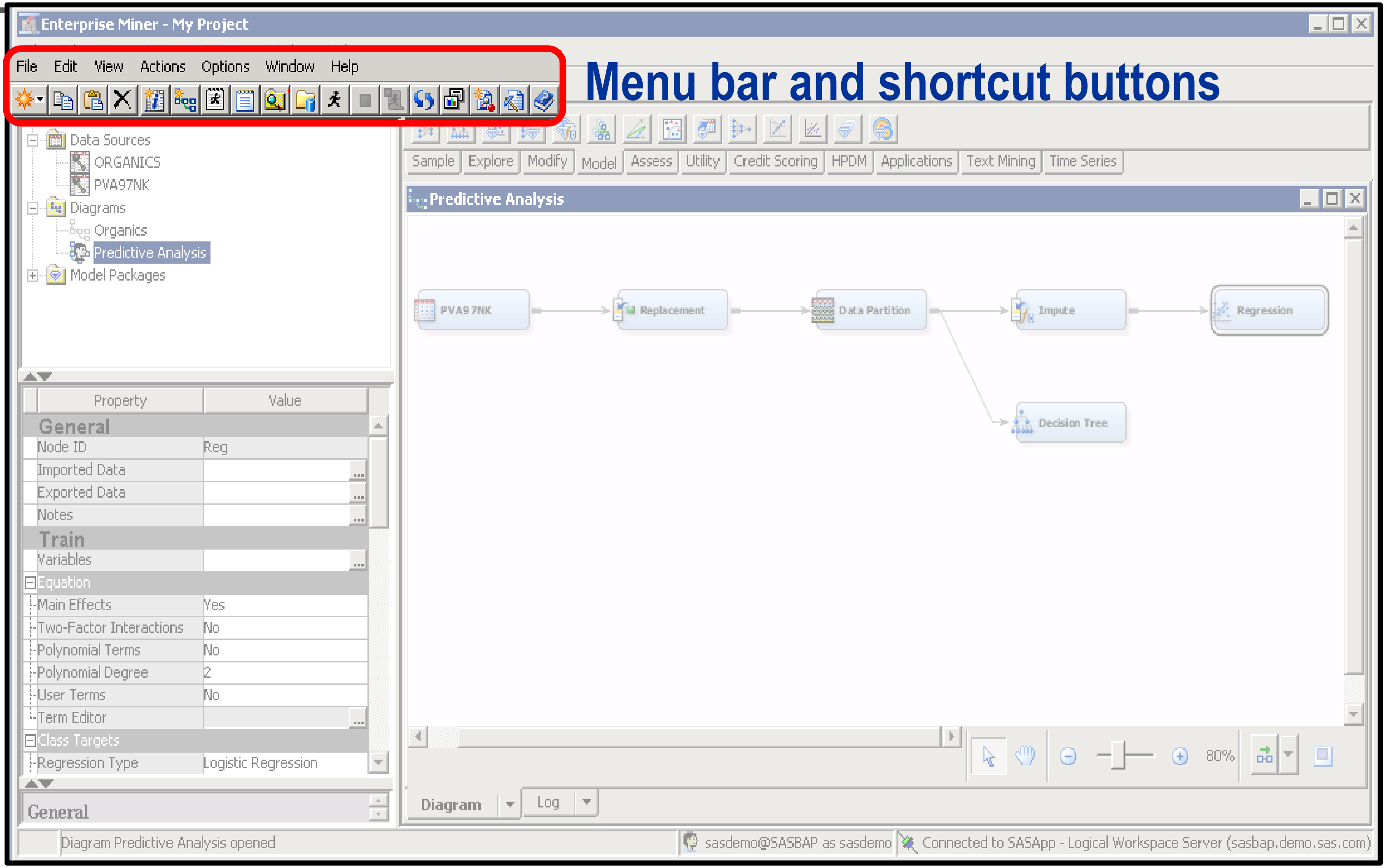
Outline

- Describe the basic navigation of SAS Enterprise Miner.
- Creating project, library (for data access) and diagram (for analysis) in SAS EM

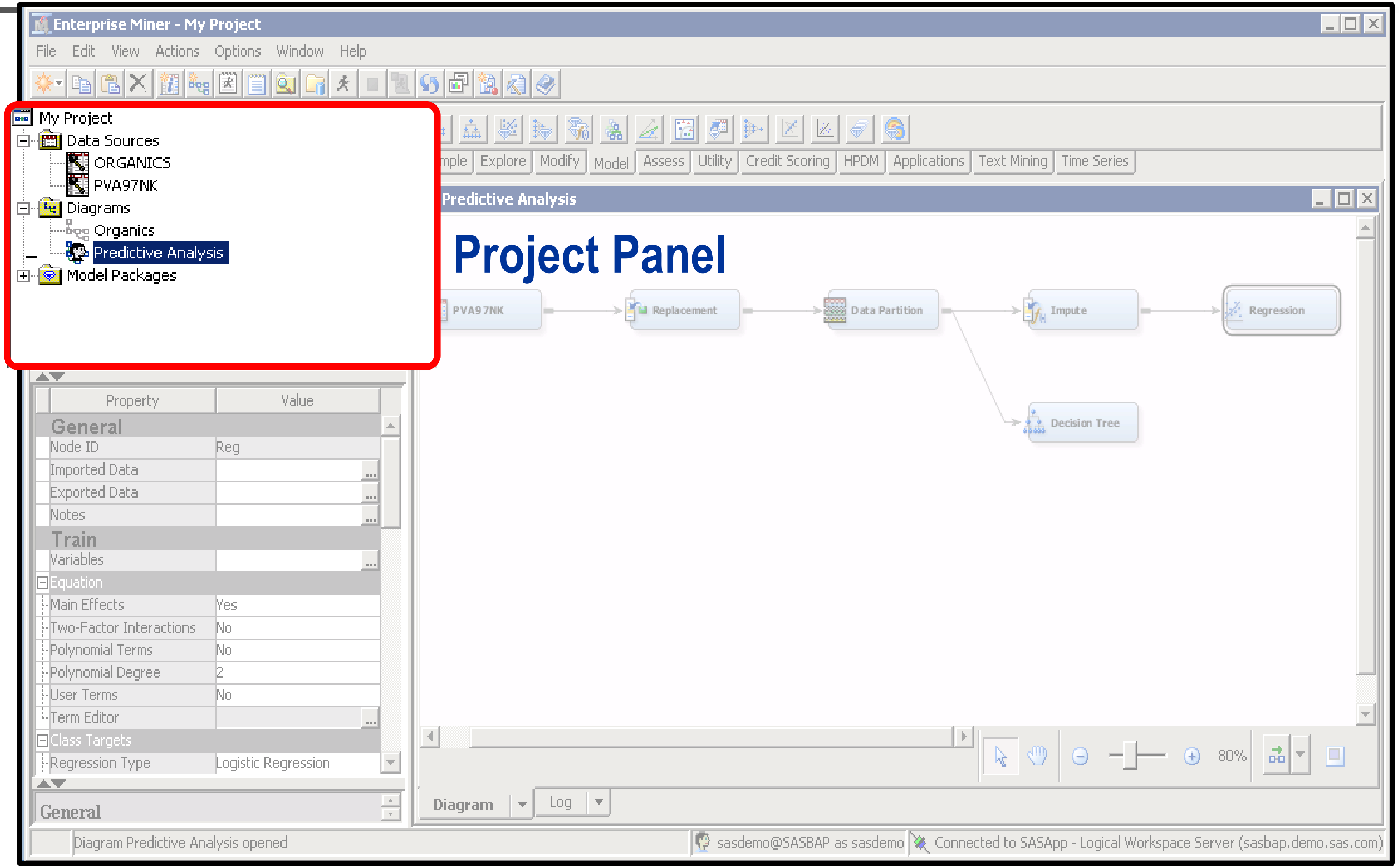
SAS Enterprise Miner



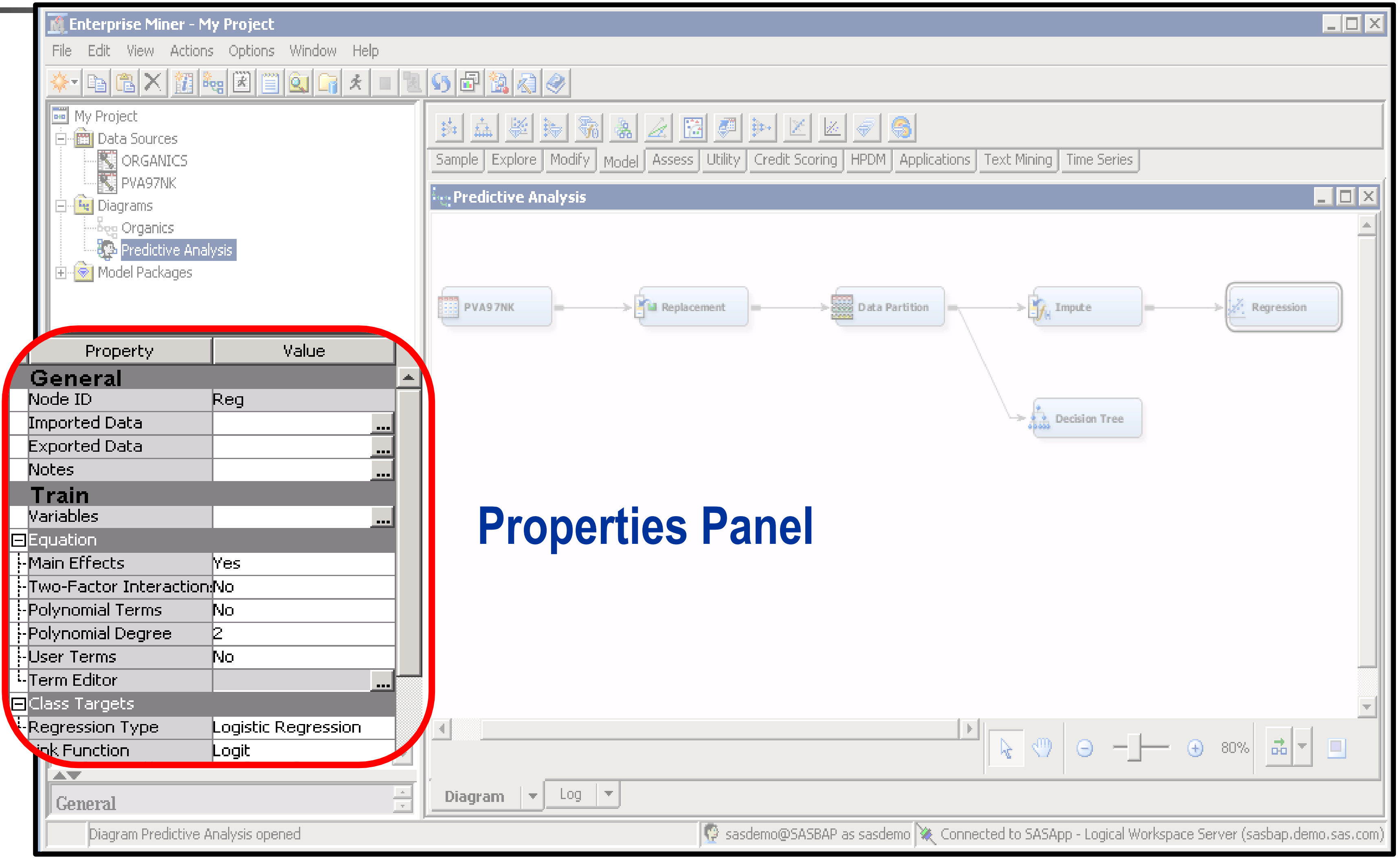
SAS Enterprise Miner – Interface Tour



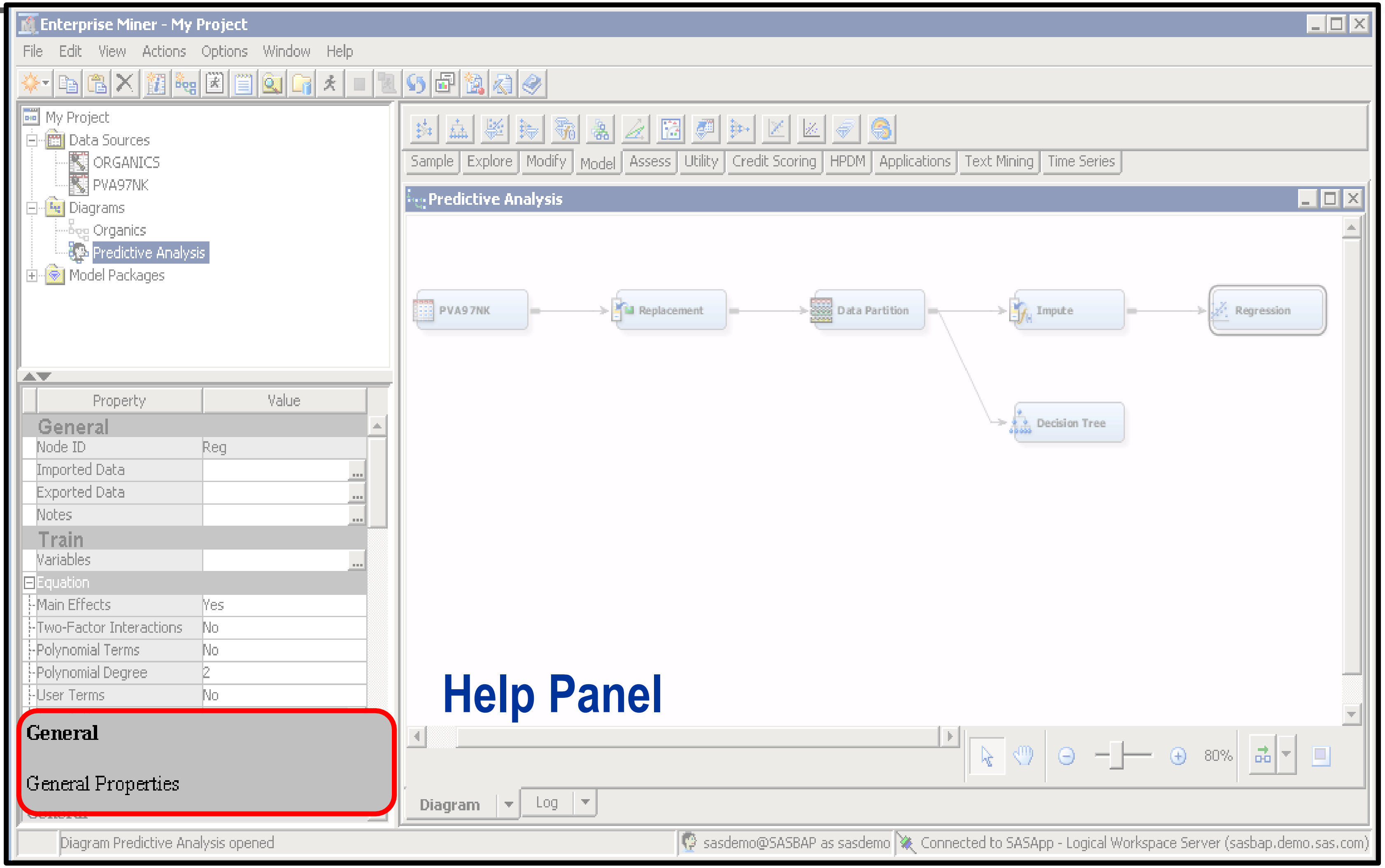
SAS Enterprise Miner – Interface Tour



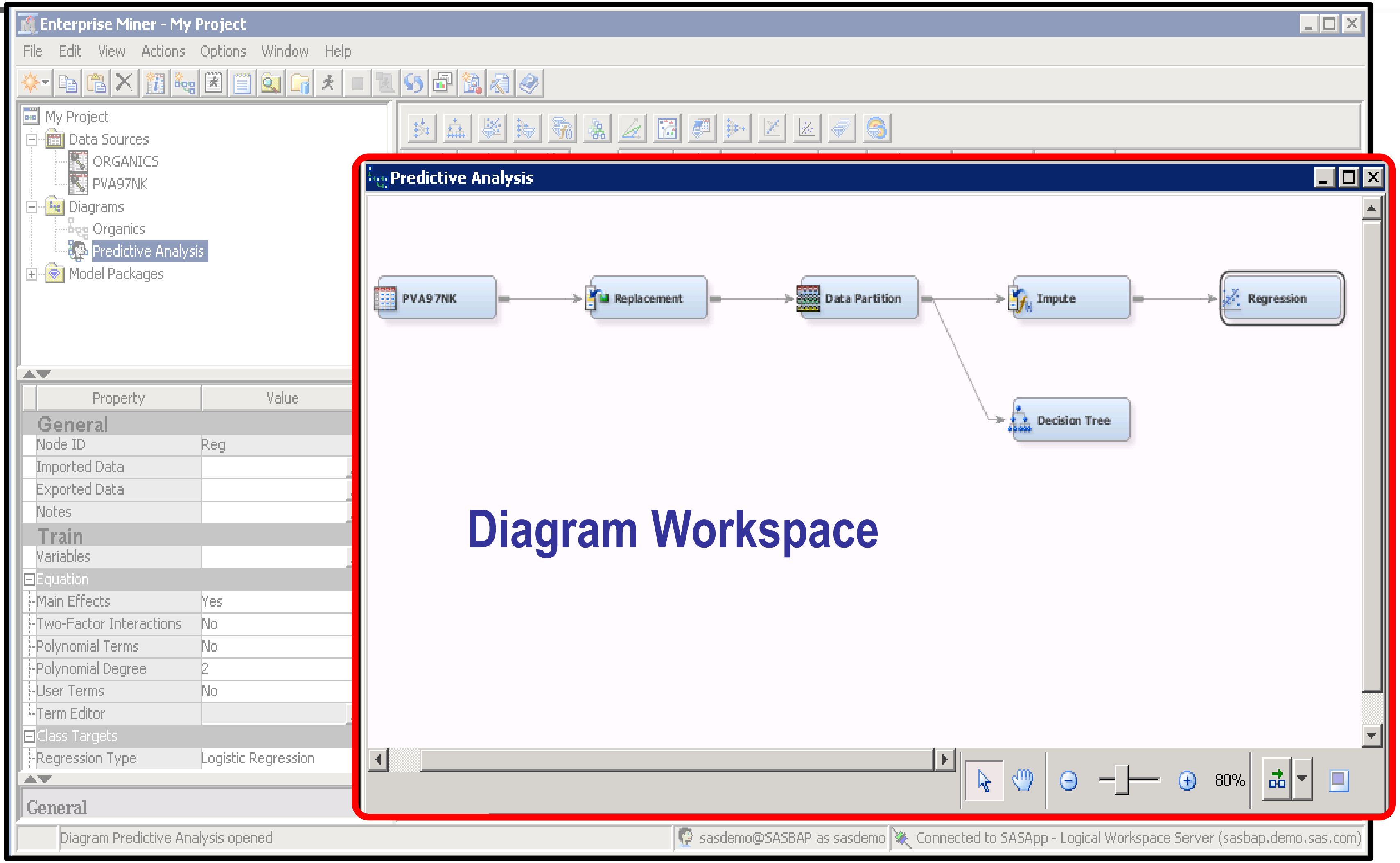
SAS Enterprise Miner – Interface Tour



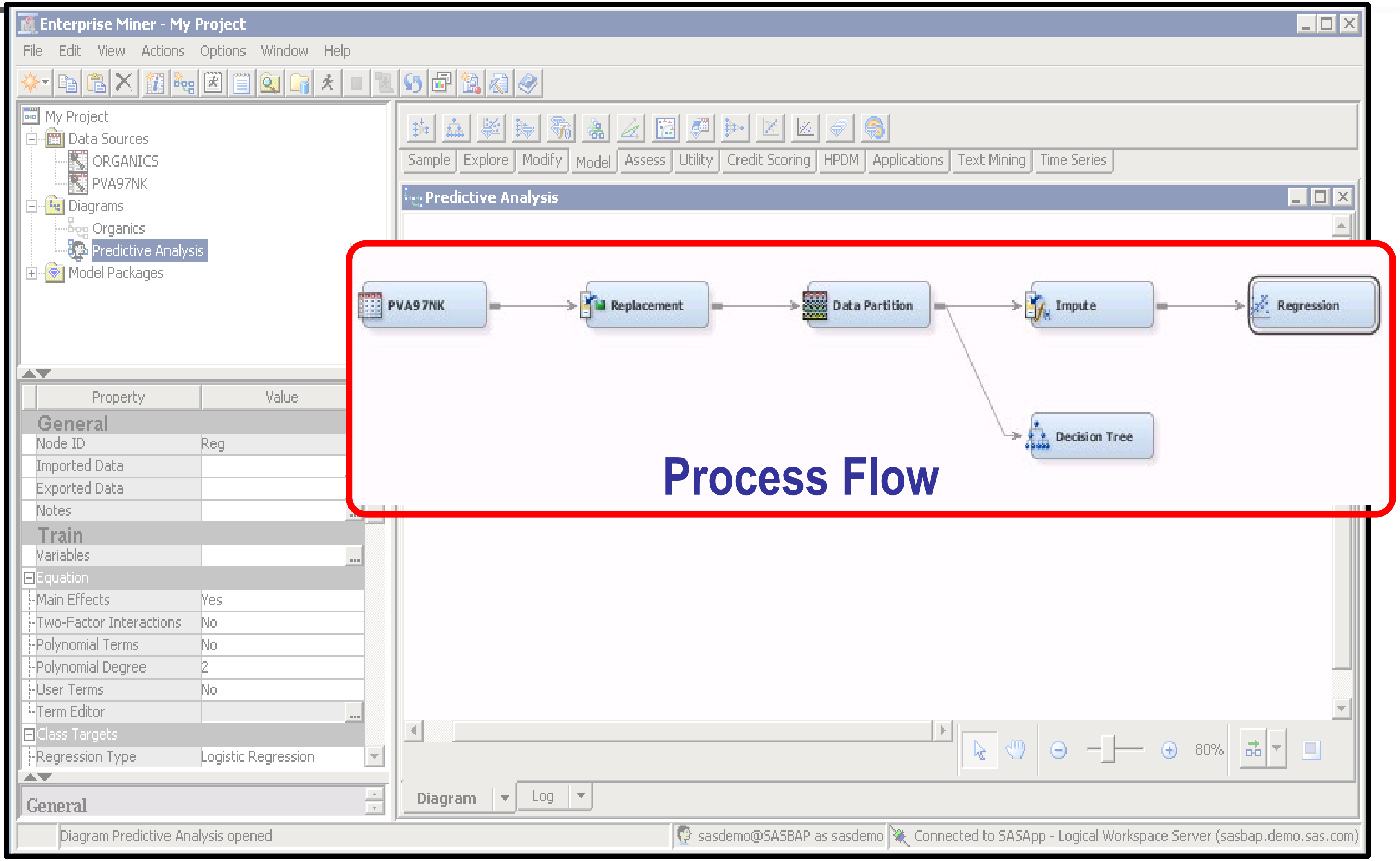
SAS Enterprise Miner – Interface Tour



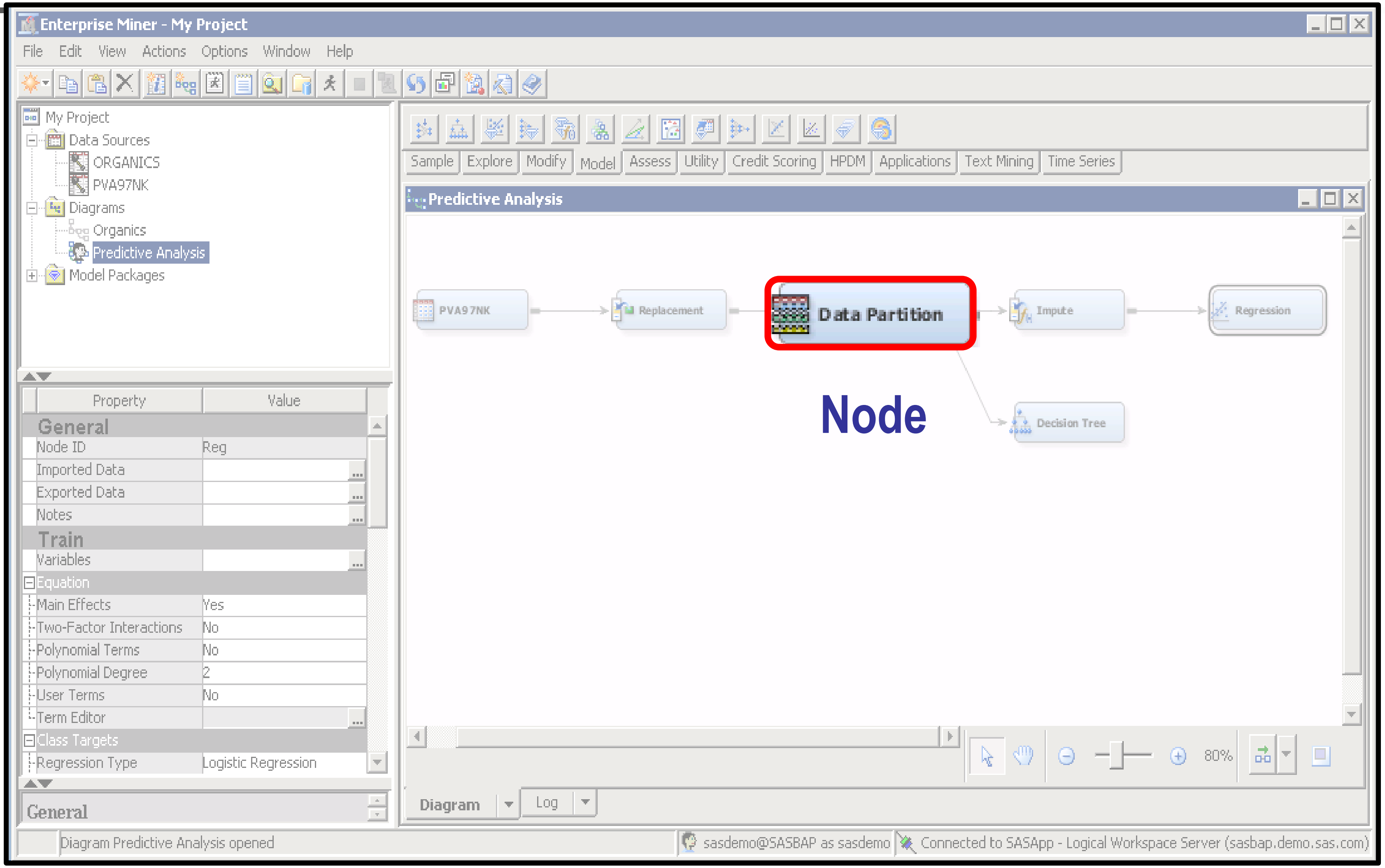
SAS Enterprise Miner – Interface Tour



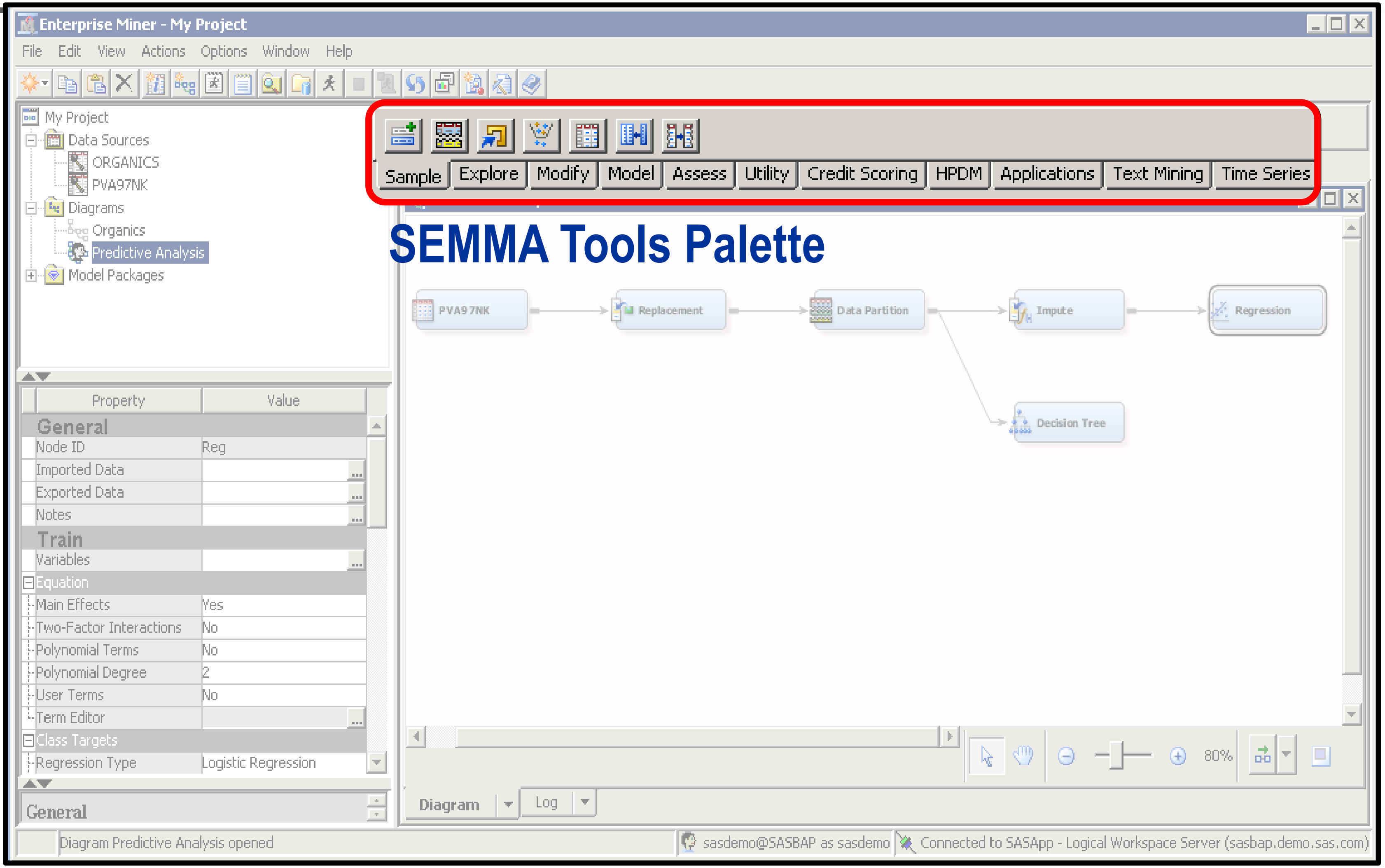
SAS Enterprise Miner – Interface Tour



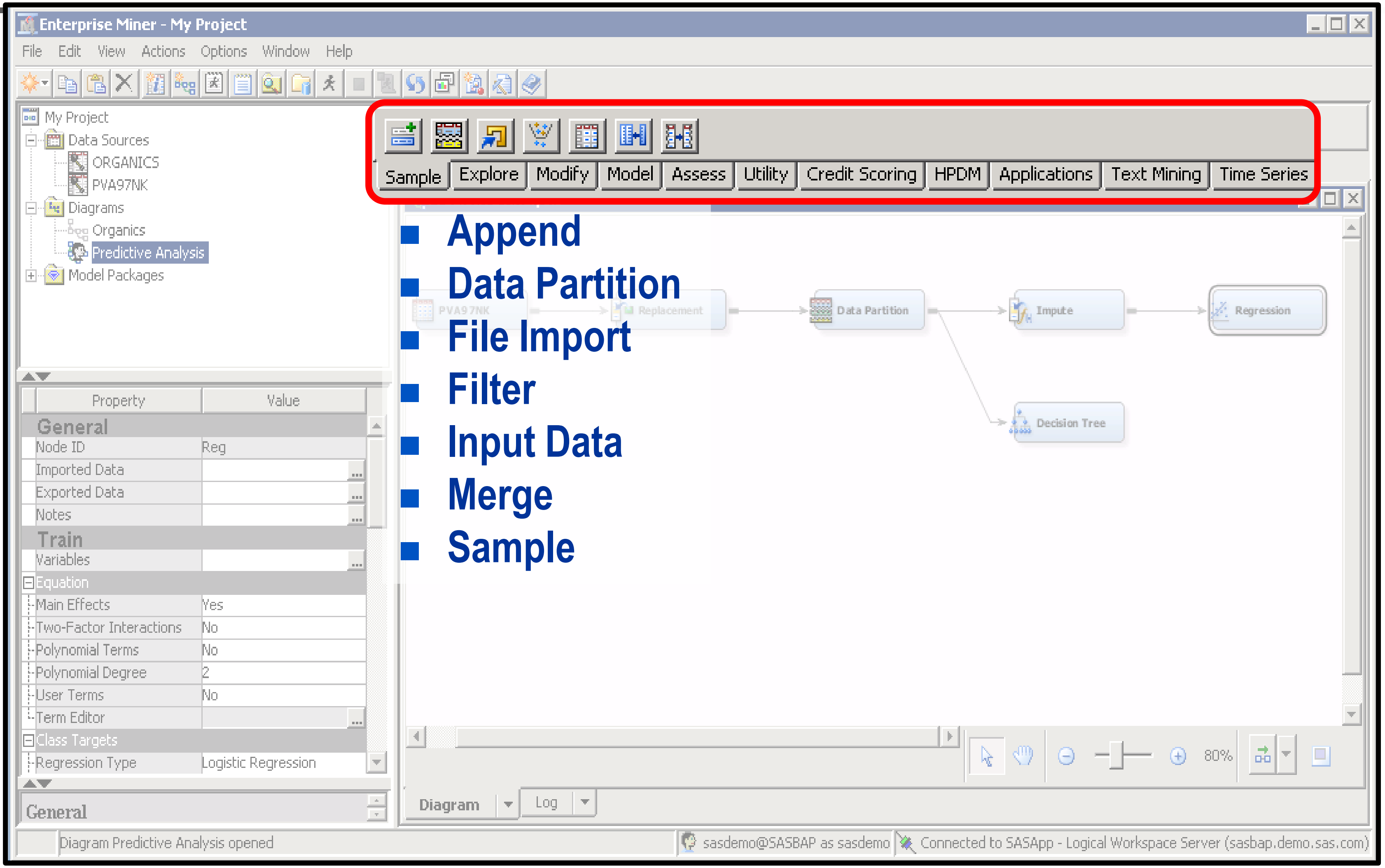
SAS Enterprise Miner – Interface Tour



SAS Enterprise Miner – Interface Tour

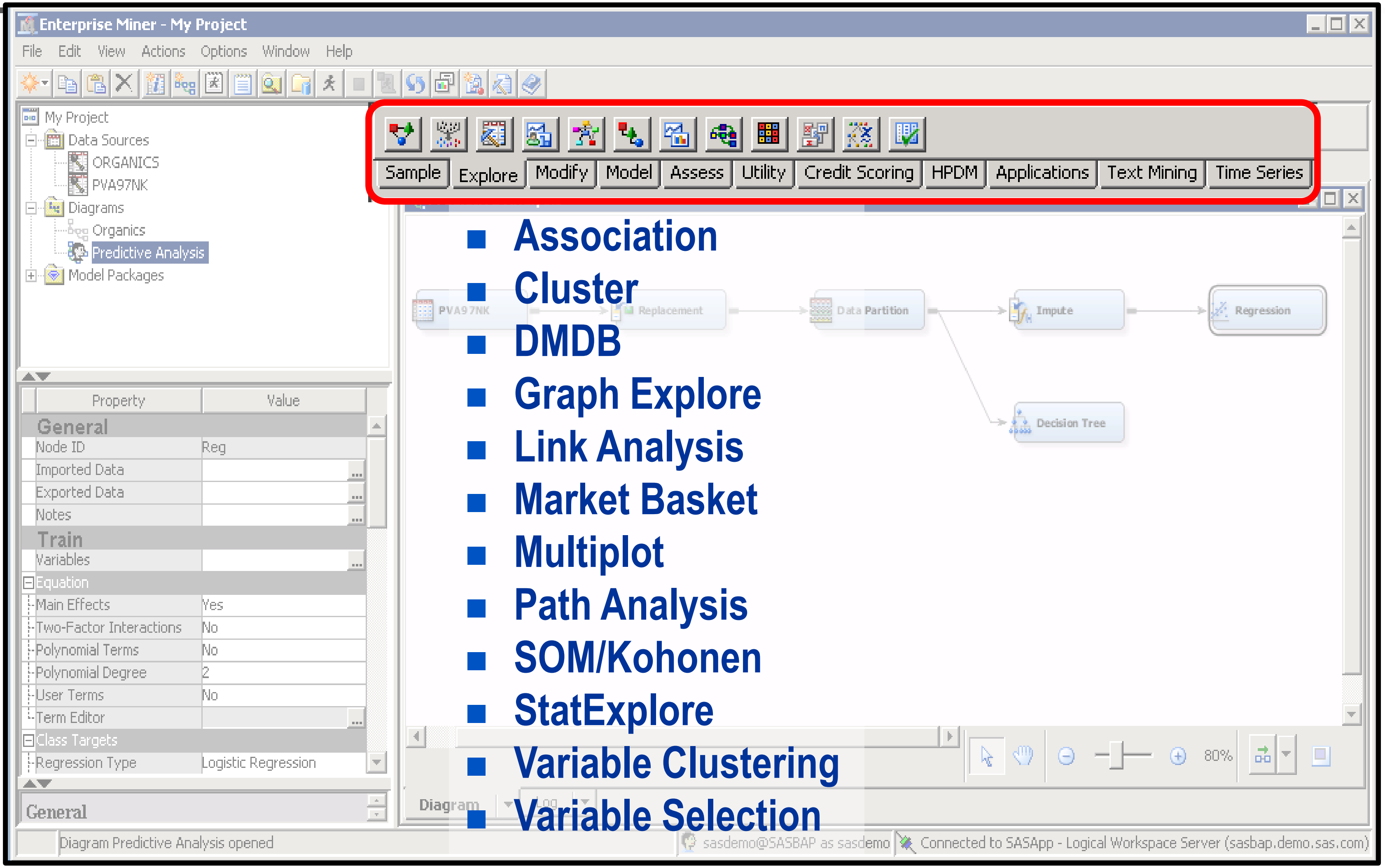


SEMMA – Sample Tab



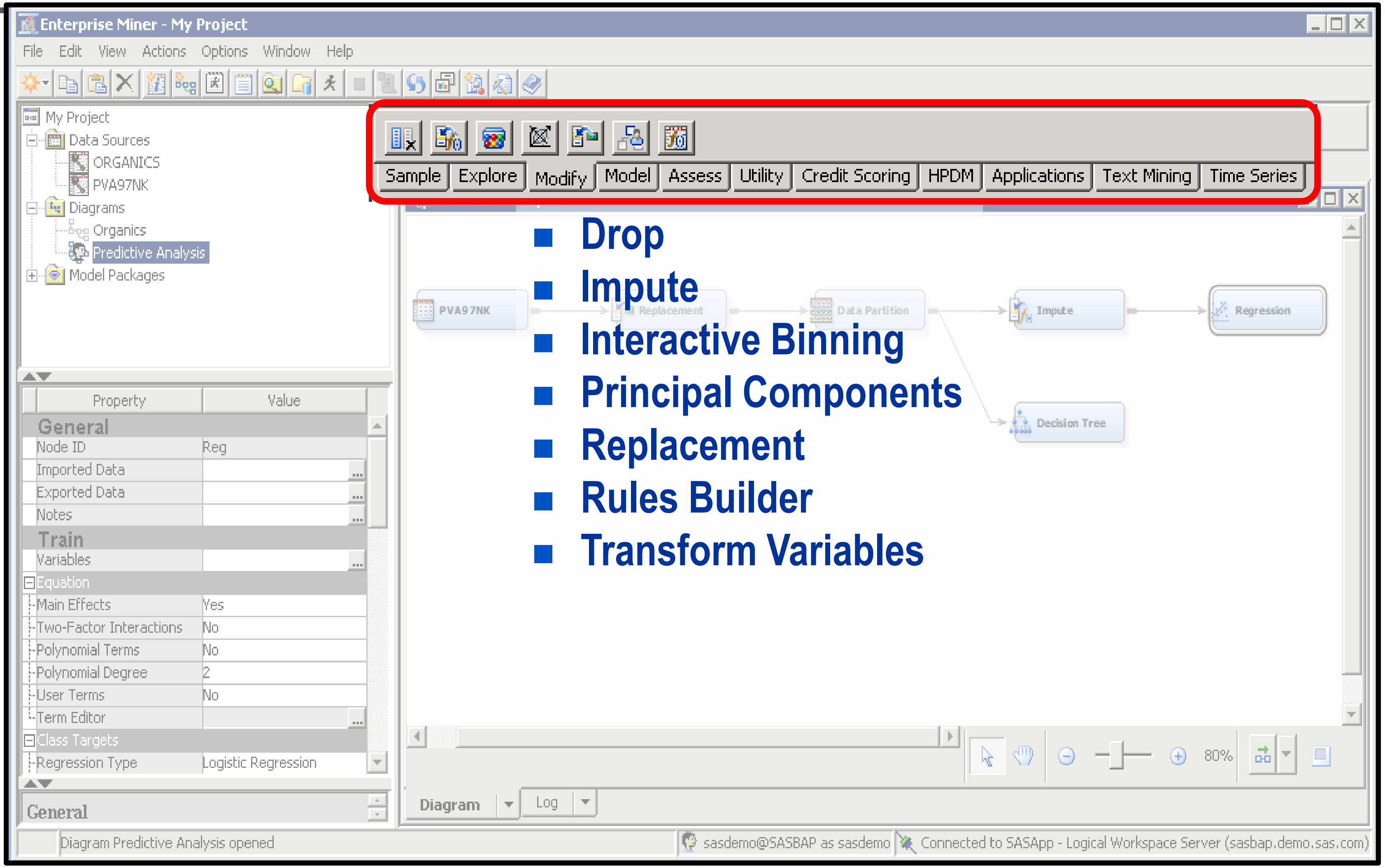
- Append
- Data Partition
- File Import
- Filter
- Input Data
- Merge
- Sample

SEMMA – Explore Tab



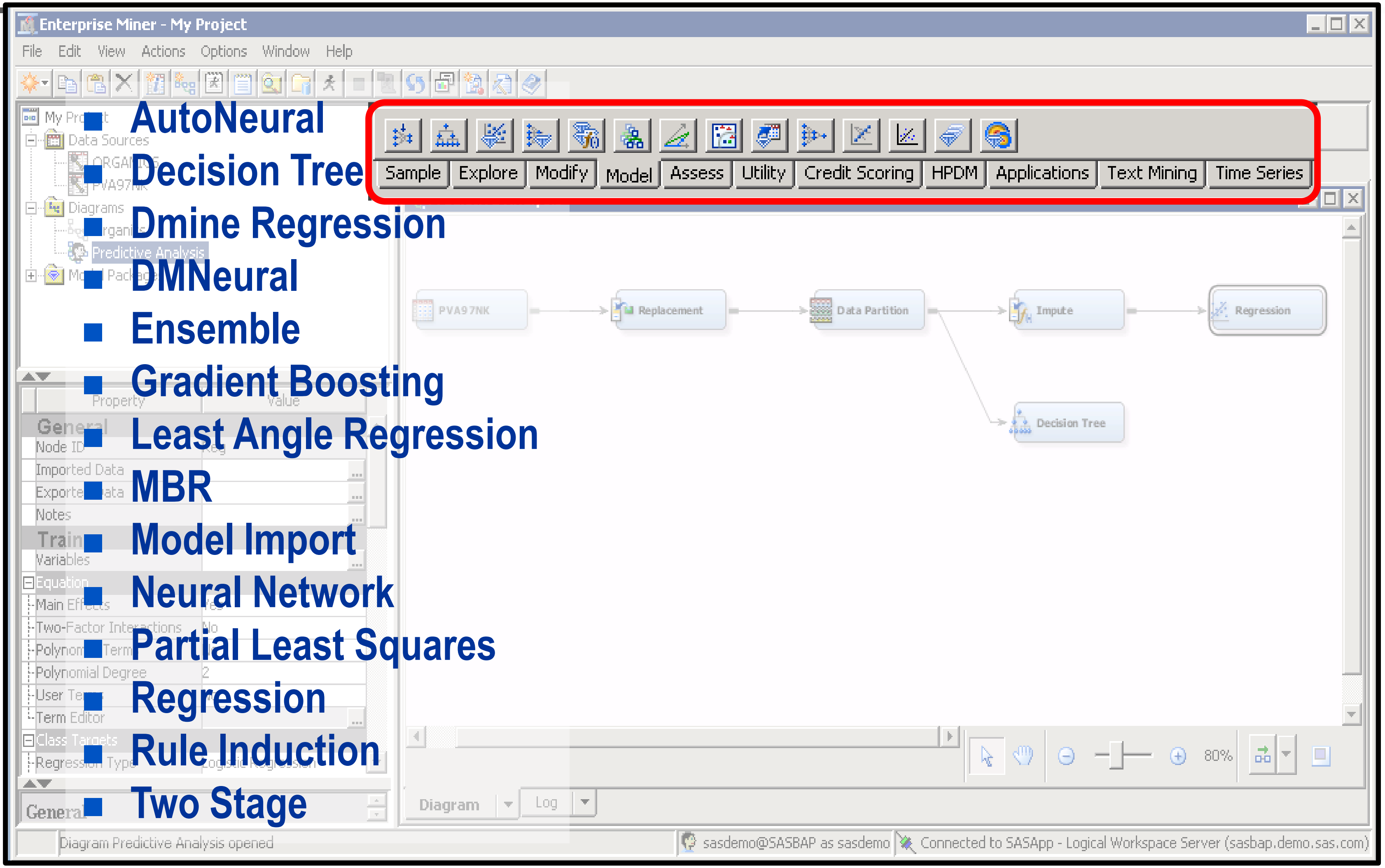
- Association
- Cluster
- DMDB
- Graph Explore
- Link Analysis
- Market Basket
- Multiplot
- Path Analysis
- SOM/Kohonen
- StatExplore
- Variable Clustering
- Variable Selection

SEMMA – Modify Tab



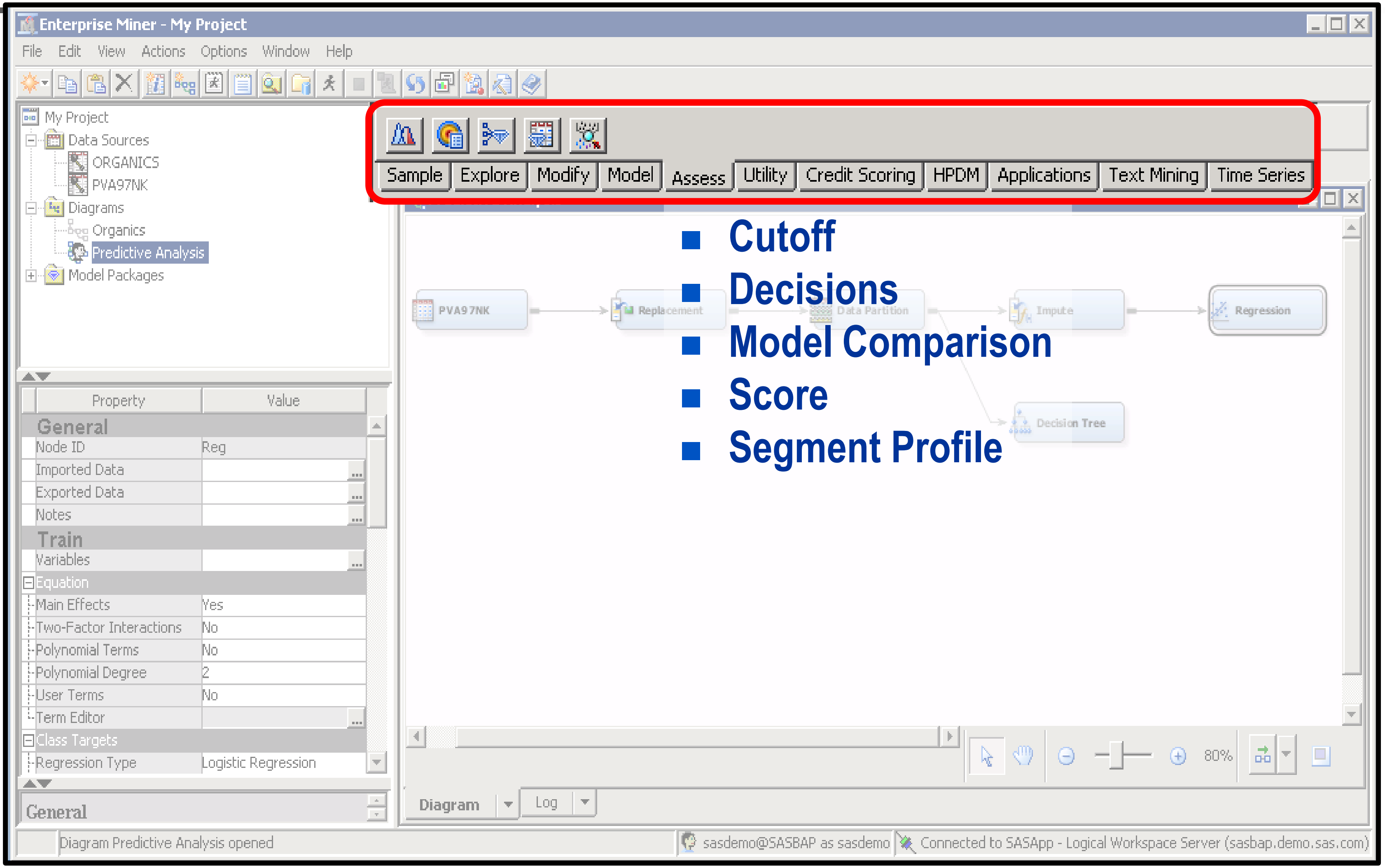
- Drop
- Impute
- Interactive Binning
- Principal Components
- Replacement
- Rules Builder
- Transform Variables

SEMMA – Model Tab



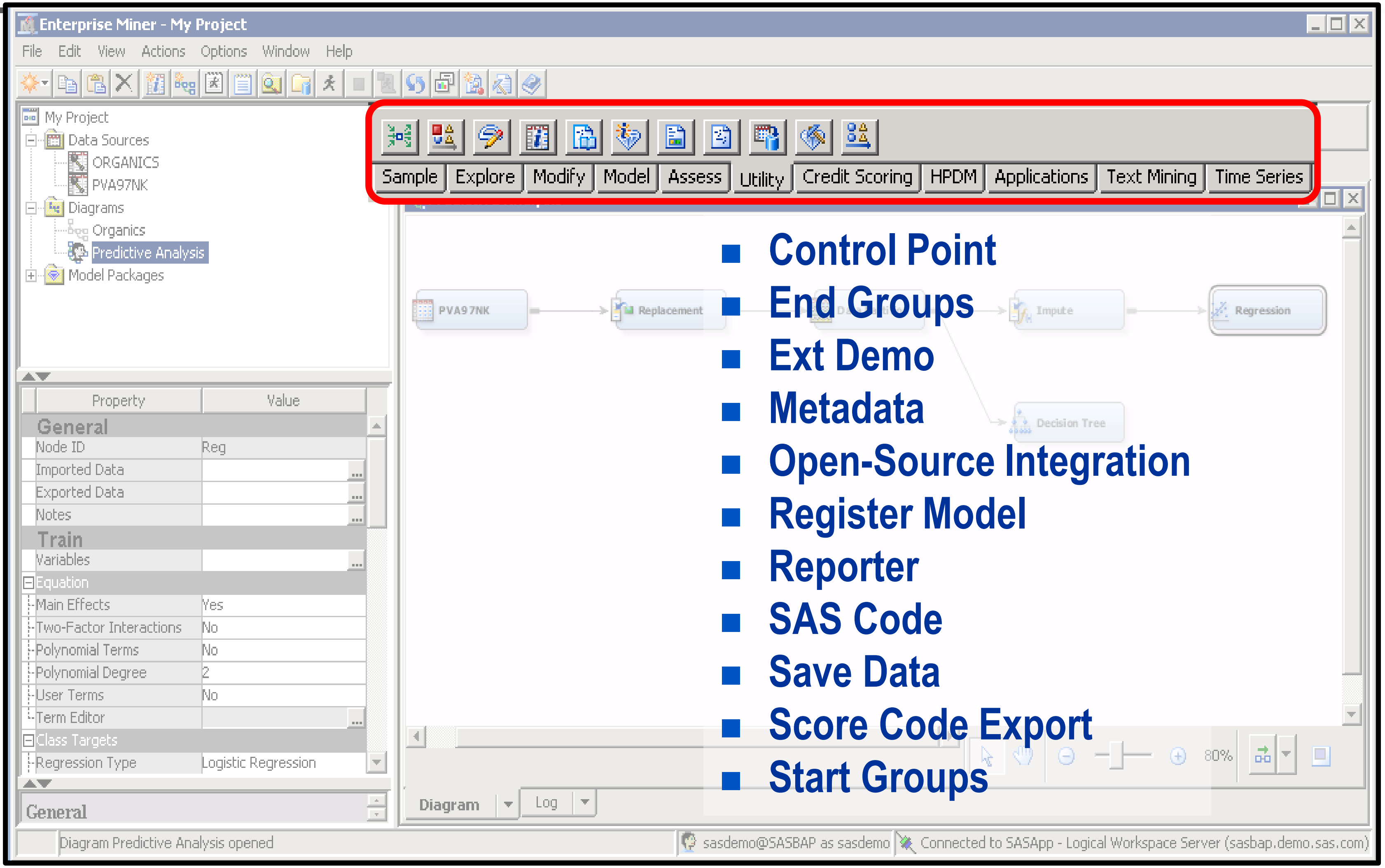
- AutoNeural
- Decision Tree
- Dmine Regression
- DMNeural
- Ensemble
- Gradient Boosting
- Least Angle Regression
- MBR
- Model Import
- Neural Network
- Partial Least Squares
- Regression
- Rule Induction
- Two Stage

SEMMA – Assess Tab



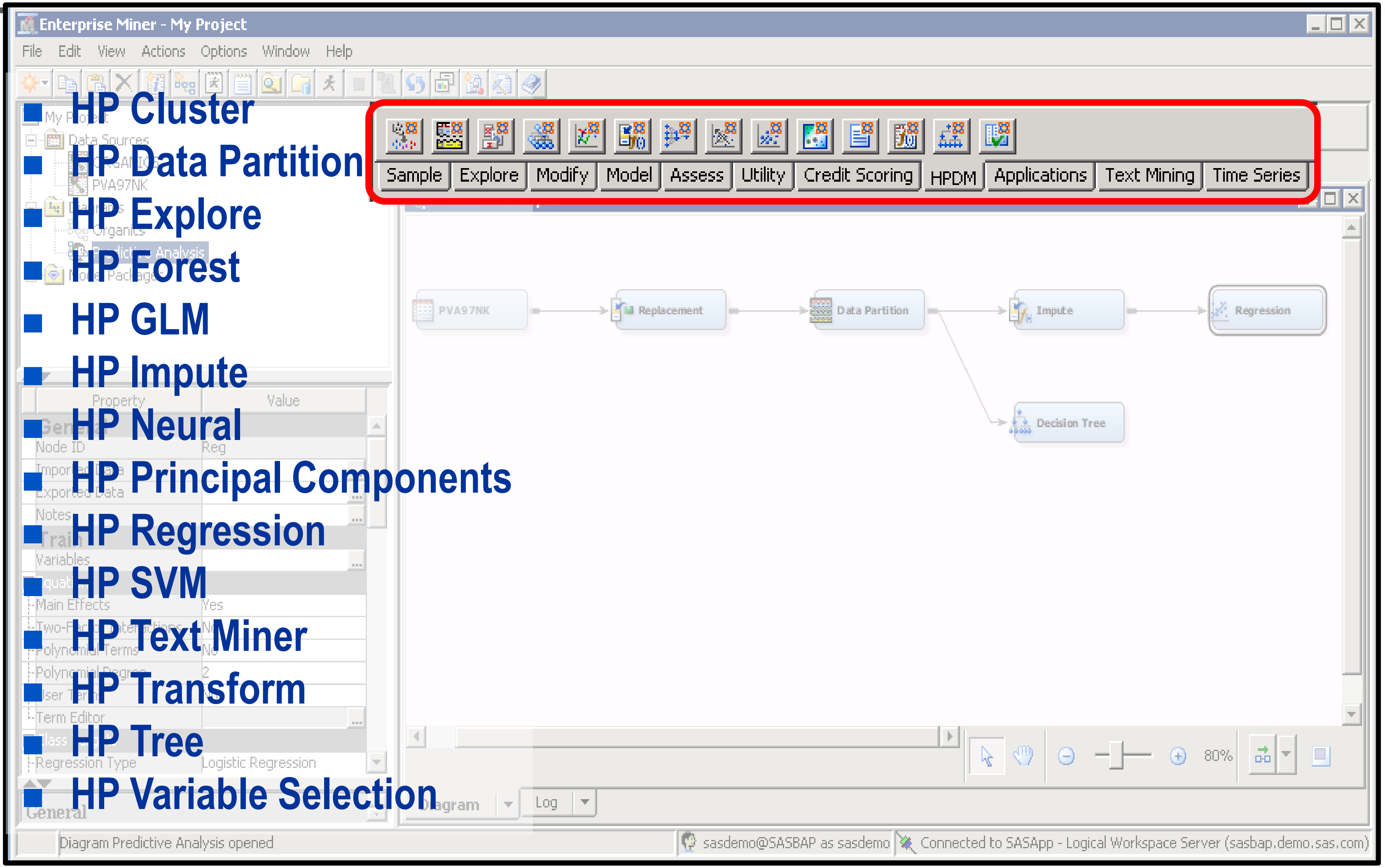
- Cutoff
- Decisions
- Model Comparison
- Score
- Segment Profile

Beyond SEMMA – Utility Tab



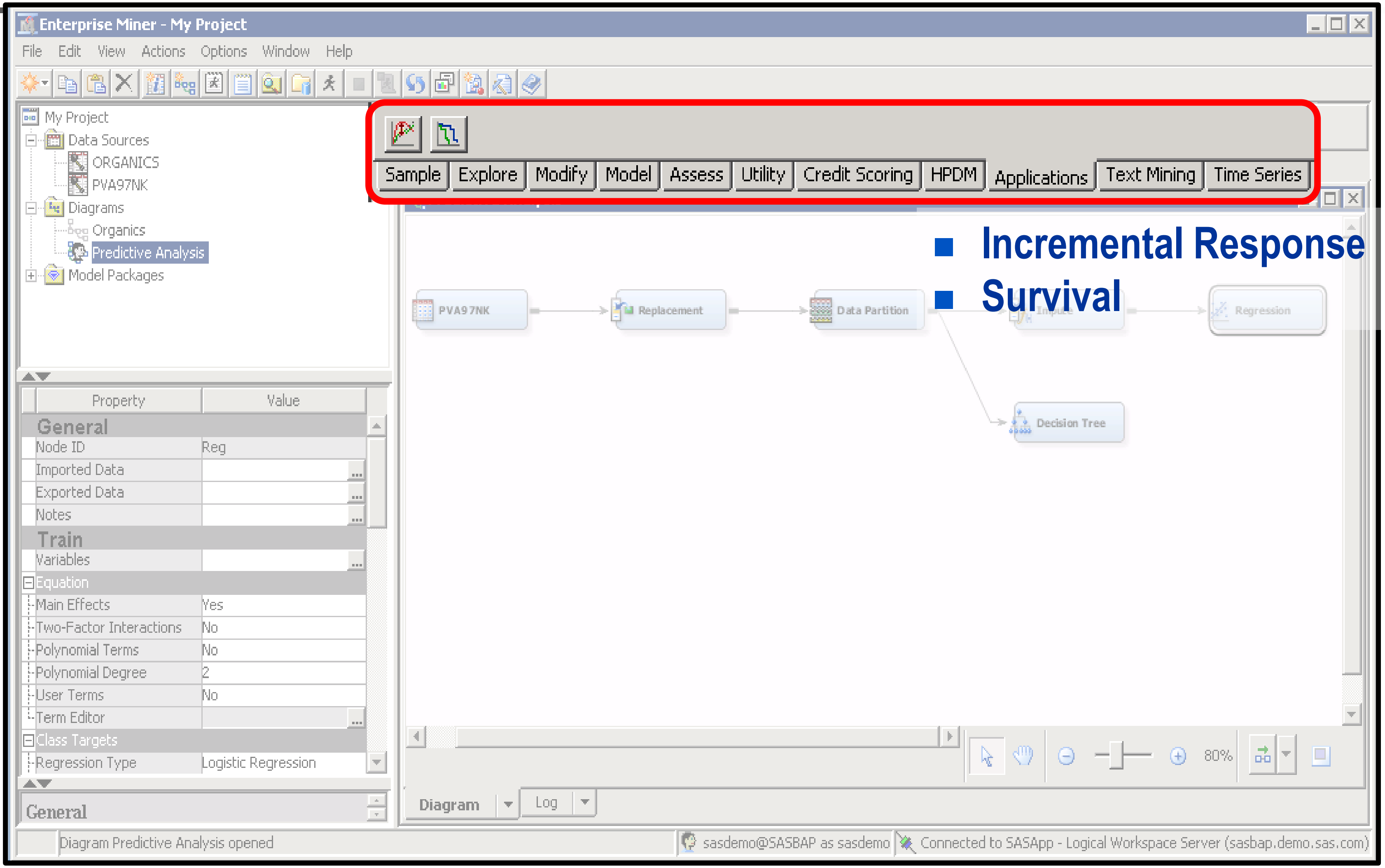
- Control Point
- End Groups
- Ext Demo
- Metadata
- Open-Source Integration
- Register Model
- Reporter
- SAS Code
- Save Data
- Score Code Export
- Start Groups

Beyond SEMMA – HPDM Tab



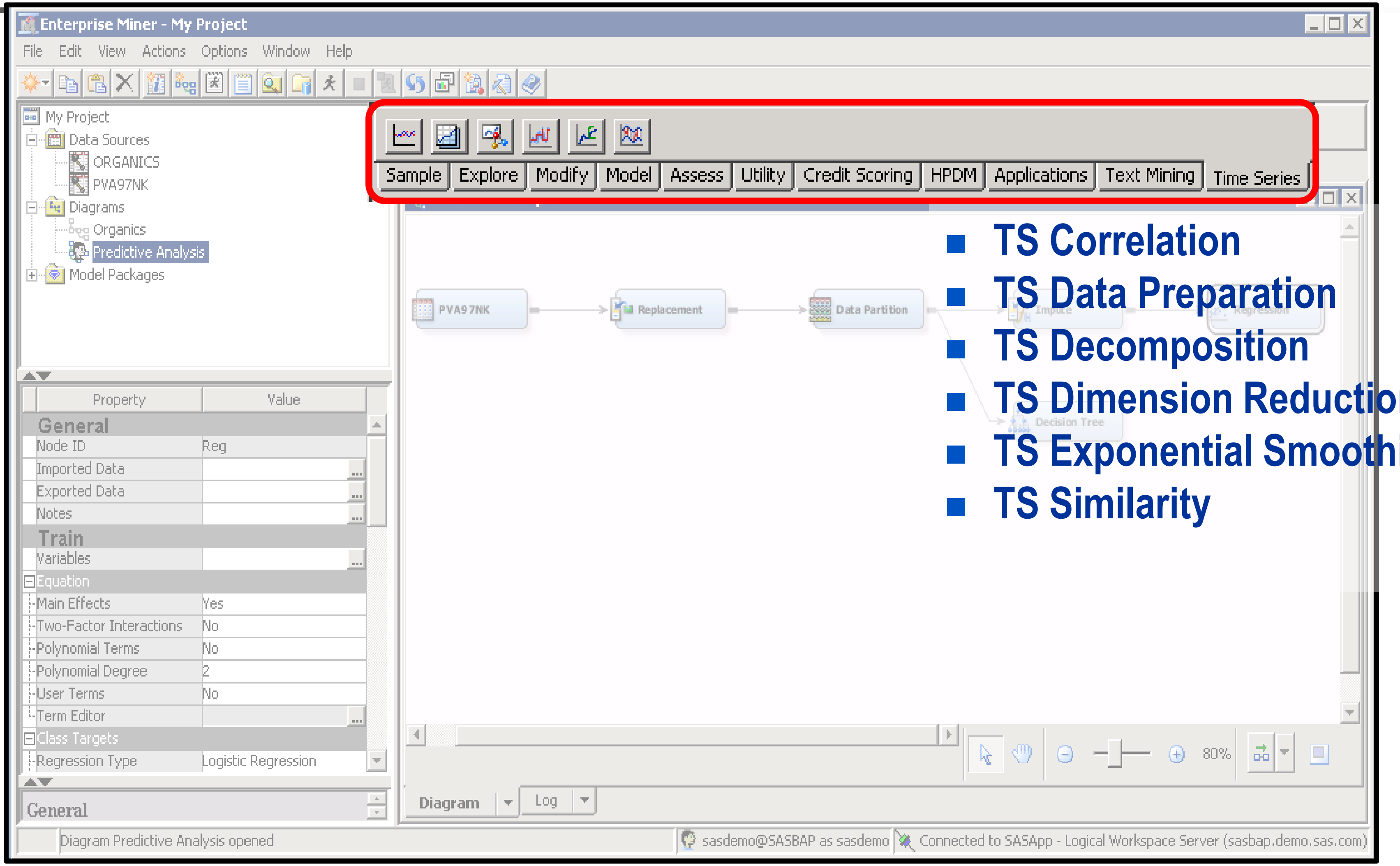
- HP Cluster
- HP Data Partition
- HP Explore
- HP Forest
- HP GLM
- HP Impute
- HP Neural
- HP Principal Components
- HP Regression
- HP SVM
- HP Text Miner
- HP Transform
- HP Tree
- HP Variable Selection

Beyond SEMMA – Applications Tab



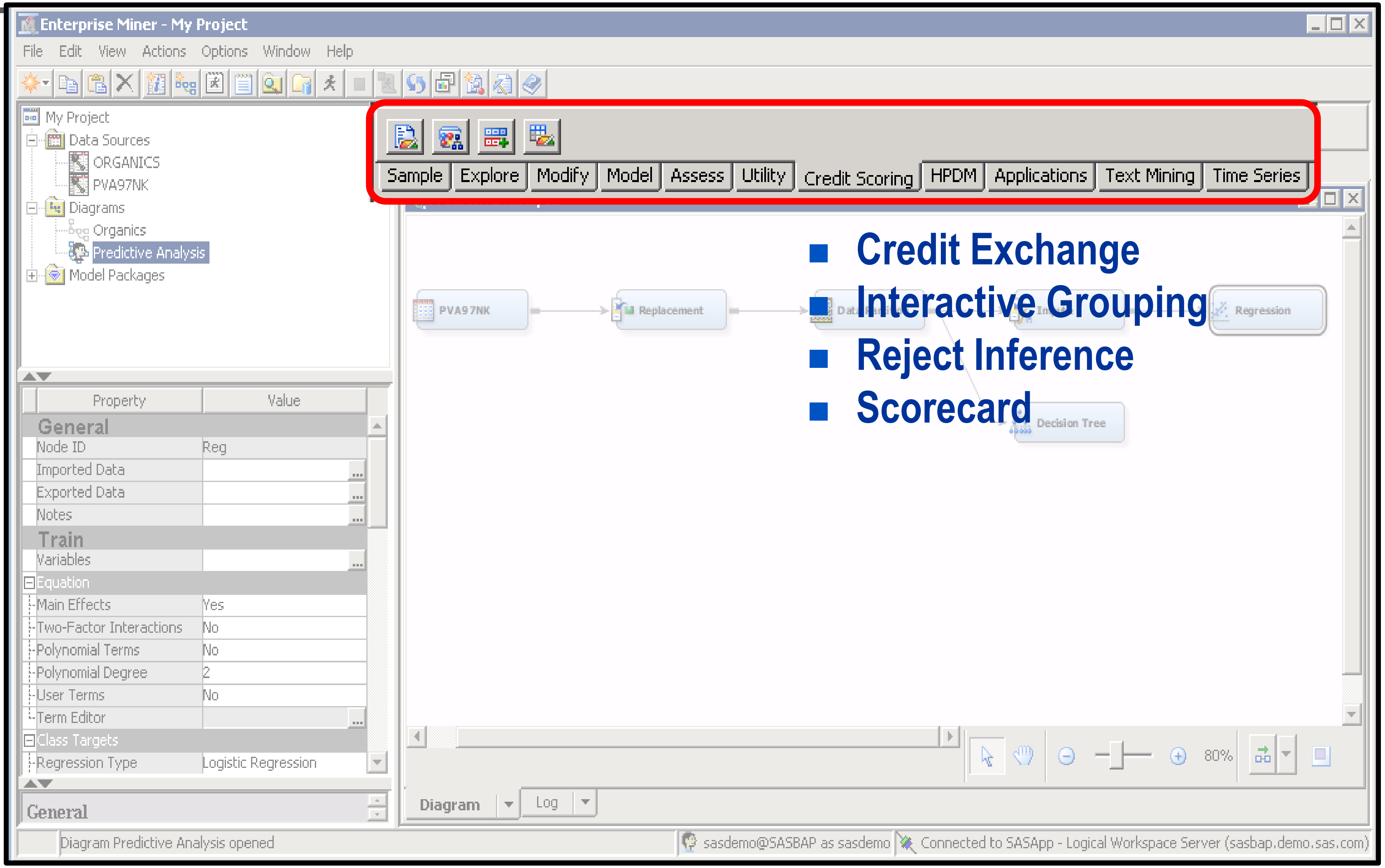
- Incremental Response
- Survival

Beyond SEMMA – Time Series Tab



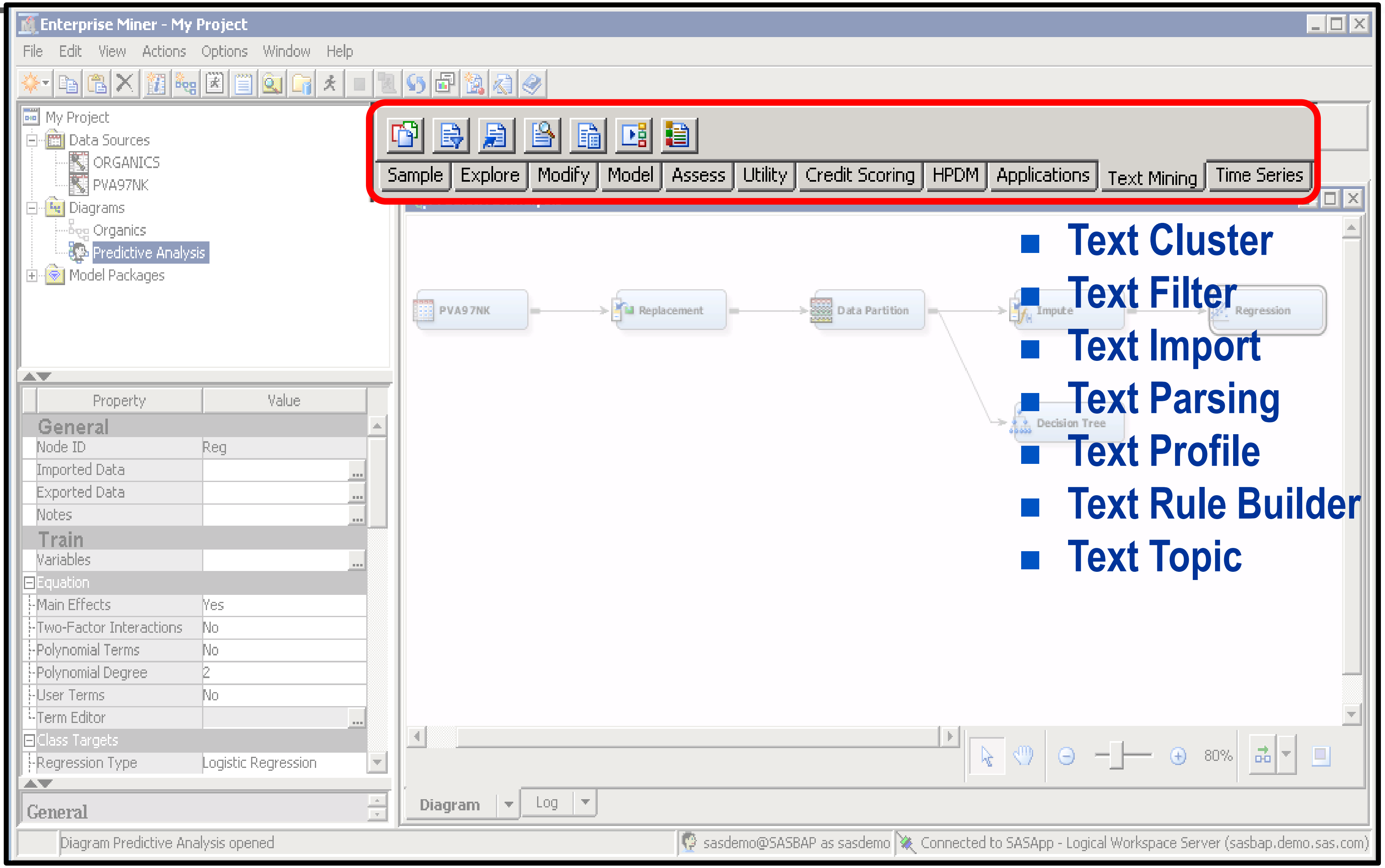
- TS Correlation
- TS Data Preparation
- TS Decomposition
- TS Dimension Reduction
- TS Exponential Smoothing
- TS Similarity

Credit Scoring Tab (Optional)



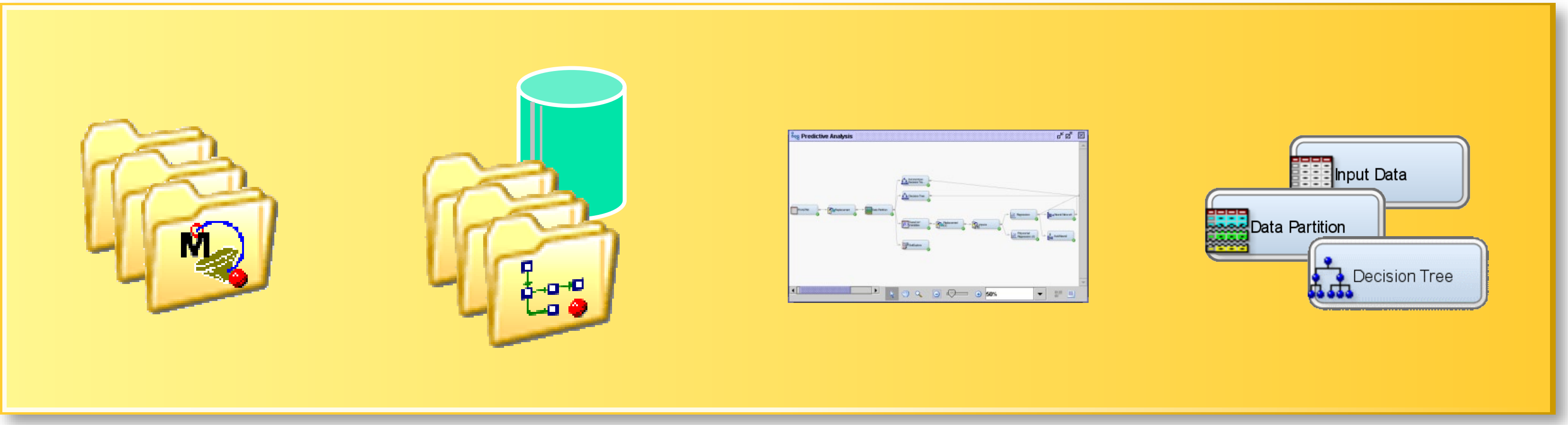
- Credit Exchange
- Interactive Grouping
- Reject Inference
- Scorecard

Text Mining Tab (Optional)



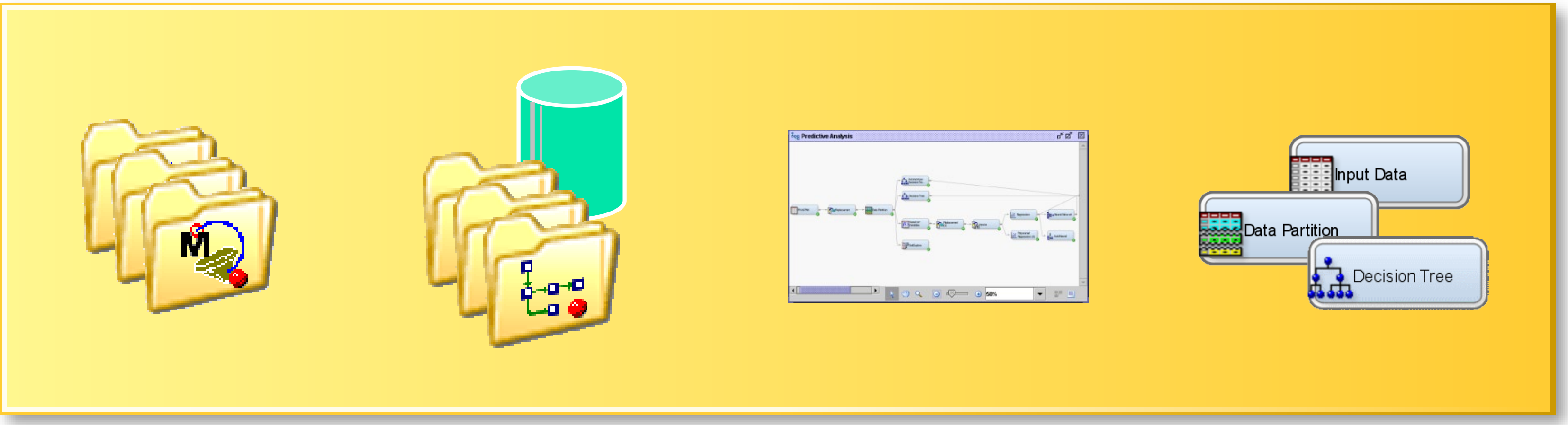
- Text Cluster
- Text Filter
- Text Import
- Text Parsing
- Text Profile
- Text Rule Builder
- Text Topic

Analysis Element Organization

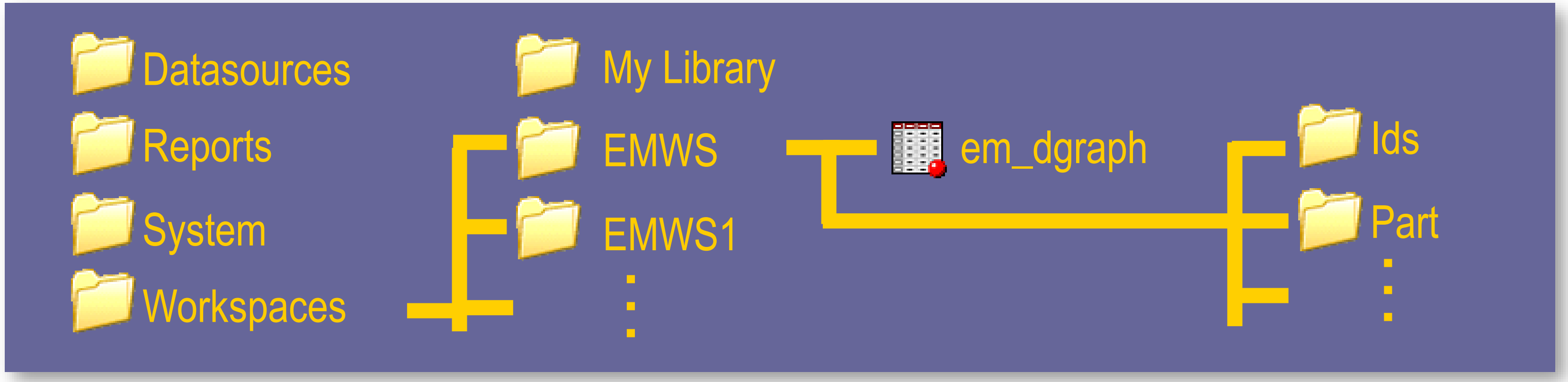


Projects **Libraries and Diagrams** **Process Flows** **Nodes**

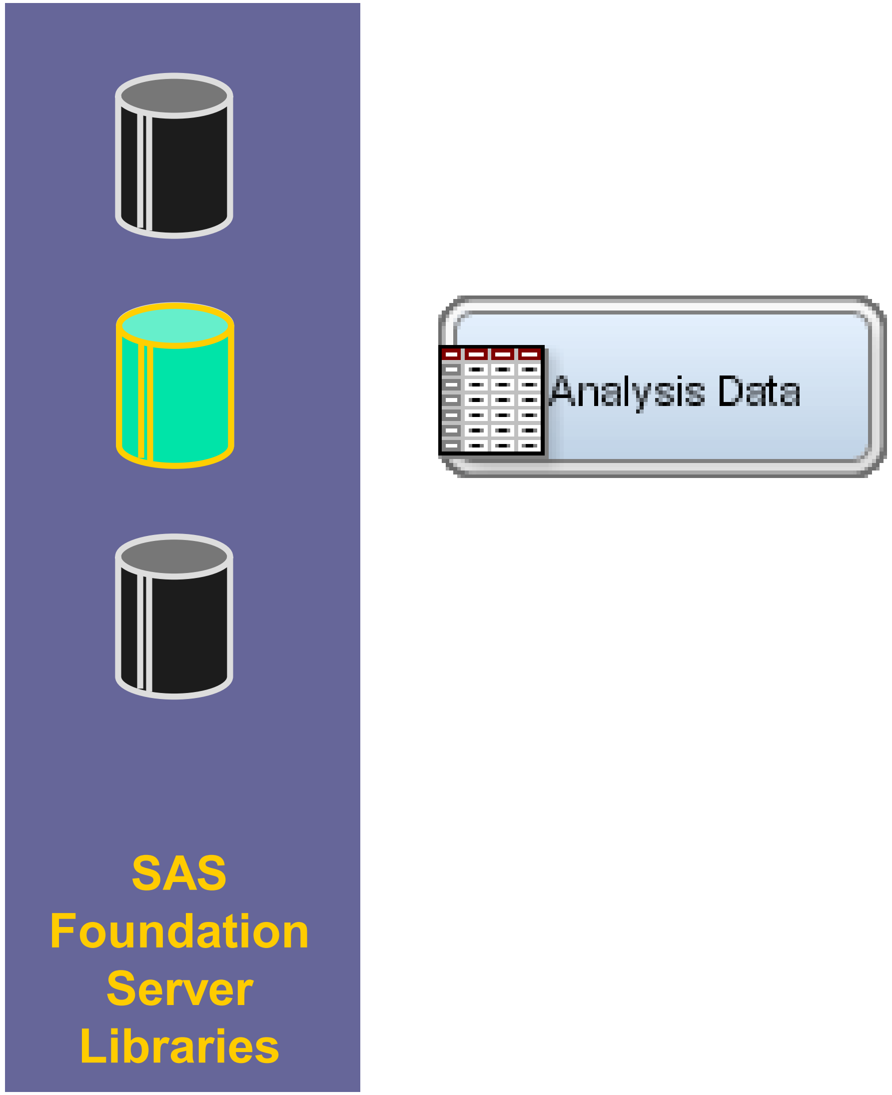
Analysis Element Organization



Projects | **Libraries and Diagrams** | **Process Flows** | **Nodes**



Defining a Data Source



- Select table
- Define variable roles
- Define measurement levels
- Define table role

SAS EM Demo

- Create a new project (save where you can access it and have enough space)
- Create a library (I will name it as *course*)
- Open the data set ***kmeans_demotr*** through the library
  - It is possible to set *specific* roles/levels of all variables in the data step creation process.
  - Best practice: *use* default selections for roles/levels of variables in data step creation process. Then, use a **Metadata** node to set *specific* roles/levels of all variables.
- Understand nature of your data via multiple methods:
  - Under data sources, right-click data table and select explore
  - In the diagram, right-click data table and select edit variables > then select variables and explore
  - Using nodes in **Explore tab** such as DMDB, Graph Explore, Multiplot and StatExplore

SAS EM Demo

Dr. Goutam Chakraborty

Outline

- Creating project, library (for data access) and diagram (for analysis) in SAS EM

SAS EM Demo

- Create a new project (save where you can access it and have enough space)
- Create a library (I will name it as *course*)
- Open the data set ***kmeans_demotr*** through the library
  - It is possible to set *specific* roles/levels of all variables in the data step creation process.
  - Best practice: *use* default selections for roles/levels of variables in data step creation process. Then, use a **Metadata** node to set *specific* roles/levels of all variables.
- Understand nature of your data via multiple methods:
  - Under data sources, right-click data table and select explore
  - In the diagram, right-click data table and select edit variables > then select variables and explore
  - Using nodes in **Explore tab** such as DMDB, Graph Explore, Multiplot and StatExplore

Transformations Before Clustering

Dr. Goutam Chakraborty

Outline

- Understand the variables and the business issues in the catalog company data set.
- Use SAS Enterprise Miner for checking distributions of variables (*bases*) and applying appropriate transformations.

A Few Cautions Before Beginning

- Most of the caveats mentioned in discussing hierarchical clustering also apply to *k*-means clustering. These include the following:
  - Selection of relevant clustering variables
  - Preprocessing of data to handle skewed distributions, outliers, missing values, and different measurement scales
  - Interpreting cluster profiles first using *bases* and then using descriptors

Business Problem and Data Description

- ABC is a supplier of identification products serving 90,000+ customers in the U.S.
- ABC wants to segment their customers based on their past and future expected transaction patterns with ABC, as well as selected firmographic variables.
  - ABC wants to consider between *2-10 segments*.
- ABC wants to profile and understand the segments using the *bases*.
- ABC also wants to profile and validate the segments using the *descriptors*.

SAS data set name: **kmeans_demoTR**

Base Variables

- **Lt_st_sales**: Total sales revenue from a customer.
- **Tele_rank**: ABC's internal estimate of ranking of customers based on future sales (smaller number is better).
- **Grow_dec**: ABC's internal estimate of which deciles customer falls in based on future growth potential.
- **RFM_group**: Seven categories (0-6) recency, frequency, and monetary grouping based on past year's transactions (higher number is better).
- **Hdcnt_last**: Number of employees in customer's location.
- **Industry**: Type of industry (based on two-digit SIC code) customer belongs in ten categories such as manufacturing, construction, and so on.

Descriptor and other managerially important variables

- **Lt_st_orders**: Total number of orders from a customer.
- **Divisions**: How many divisions within ABC a customer is buying from.
- **Acct_recency**: Time in months since last purchase.
- **Type_customer**: Four categories (platinum, gold, growable, and unspecified) of customers.
- **Reseller**: Whether the account is a reseller of ABC's products.
- **Zone**: Five categories of customer's primary location in the US (Western, Central, North, and NE South, and SE, other).
- **Credit_risk**: ABC's internal estimate of customer's credit risk (five categories).

Plan of Analysis

1. Explore this data set using SAS Enterprise Miner. In particular, look at the distributions of *base* variables.
2. Use transformation as appropriate on *base* variables.
3. Run *k*-means using SAS Enterprise Miner.
4. Interpret results from *k*-means.

Why Do Base Variable Transformation before Clustering?

- To give *equal importance* to each variable in influencing cluster results
- To reduce *Skewness and Kurtosis* to a manageable number
- Ideally, we would like variable distribution to be close to Normal (if that's not possible, at least ….)

Types of Base Variable Transformation

For Numeric Variables:

- Scale transformation
  - Range or Centering transformation that does not change shape of the distribution
- Shape transformation
  - Power series and other transformations (such as double-standardization) that change both scale and shape
  - Examples are square, square root, inverse, log, and so on
- *Numeric to Categorical* transformation
  - Quantile, Bucket, Optimal Binning, and so on

For Categorical Variables:

  - Combine very rare classes into “other” class
  - Convert to numeric via WOE method

Checking Distributions and Handling Transformations

- This demonstration illustrates using SAS Enterprise Miner to get a feel for data, checking distributions, and handling transformations.

Summary of Checking Distributions and Handling Transformations

- Of the six base variables, there are three numeric variables and three categorical variables.
- The three categorical variables do not seem to have *very rare* classes.
- Of the three numeric variables, **HDCNT_LAST** and **lt_st_sales** show *large, right skew*.
  - Max. Normal method indicated log transformation for these two variables.
- Max. Normal method indicated square root transformation for the variable **tele_rank**.

Demo of k-Means

Dr. Goutam Chakraborty

Outline

- Running k-Means and interpreting results

Plan of Analysis

1. Explore this data set using SAS Enterprise Miner. In particular, look at the distributions of *base* variables.
2. Use transformation as appropriate on *base* variables.
3. Run *k*-means using SAS Enterprise Miner.
4. Interpret results from *k*-means.

Applying *k*-Means

- This demonstration illustrates how to run *k*-means clustering and interpret the results.

Profiling k-Means Clusters

Dr. Goutam Chakraborty

Outline

- Profile kMeans clusters with base variables
  - Instead of using transformed variables, use the raw (untransformed) base variables for ease of business understanding.
- Profile kMeans clusters with descriptor variables

Recap Profiling Clusters : The Big Questions

Several types of questions are often asked in profiling:

- How is the average member *of one cluster* different from an average member *of a different cluster*?
- How is the average member *of any cluster* different from the average member *of the entire data*?
- How does the *distribution* of a variable *within a clus*ter compare to the *distribution* of the same variable in the *entire data*?
- Which variables are *most important predictors* for **each** cluster?

Recap Profiling Clusters with Bases

- Profiling involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters.
  - For **numeric** variables, this involves
    - comparing the mean of each variable across clusters
    - comparing the mean of each variable in a cluster with the mean for the same variable for the entire data
    - comparing the distribution (histogram) of each variable in a cluster with the distribution of the same variable for the entire data
  - For **categorical** variables, this involves comparing % members in each category within a cluster with the % members in the same category for the entire data

Profiling Clusters using Base Variables

- Save/export data from SAS EM using SAS code and then Use SAS EG and ANOVA on the saved data.
  - Not demonstrated but you should try on your own
- Use **Segment Profile** node in SAS EM
  - Set roles of untransformed base variables from "rejected to input"
  - Use results from segment profile node along with means/frequencies from SAS code (see below) to tell a story about each cluster
- Use SAS Code in SAS EM to get the *means by clusters* for interval variables and *cross-tab by clusters* for categorical variables
  - Create index and report index instead of raw mean/frequencies – not demonstrated here but you should try on your own

Profiling *k*-Means Clusters

- This demonstration illustrates how to profile k-mean cluster using bases.

Summary of Cluster Profiles (Bases)

Bases	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6
Life Time Sales	$2,217	$9,409	$1,280	$1,084	$8,269	$3,595
Grow Dec	More of d10	More of d01 – d03, d05	More of d08- d10	More of d08- d10	More of d04	More of d0-d03
Tele Rank	55,989	26,529	64,252	60,470	28,234	41,981
RFM Group	Mostly Groups 2 and 1	Group 6	Group 4	Group 0	Mostly Groups 3 and 6	Group 5
Employee Count	117	208	100	97	196	144
Industry	Slightly More Education	Slightly More Manufacturing	Slightly More Services	Somewhat More Unclassified	More Manufacturing	Slightly more public admin

Profiling *k*-Means Clusters Using Descriptors (**Self Study**)

- This demonstration illustrates profiling *k*-means clusters using descriptors.
- Do it on your own. **Note**: set the value of minimum worth in segment profile node to 0.001 (it was 0.01 for base variables)

Summary of Cluster Profiles (Descriptors)

Descriptors	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6
Type Customer		More gold, growable and platinum			More growable	More growable
Acct Recency	8.8	6.1	8.2	4.7	6.6	7.6
Divisions	1.6	2.0	1.5	1.5	2.0	1.8
Lifetime Orders	7.9	29.0	4.6	3.9	27.1	12.5
Credit Risk	Slightly more New and 1001	More 1003	More New	More New	More 1003	More 1003
Zone		Slightly more North NE			Slightly more Central	

Next Steps (for You to Do on your own)

- Sort the data differently and rerun cluster analysis to check for order effect.
  - This is one way to force the algorithm to use a very different set of starting seeds.
- Use different transformations on the base variables.
- Trim (or, Winsorize) outliers/atypical observations.
- Use a different method (you used Average) such as Ward's or Centroid method for the first stage in the clustering algorithm.
- Force a different number of cluster solutions (by switching from automatic to user specify in SAS Enterprise Miner and then specifying the number of clusters) and evaluate those solutions.