# Documentation for SAAS-Enterprise Miner

*Talend Data Prep:*

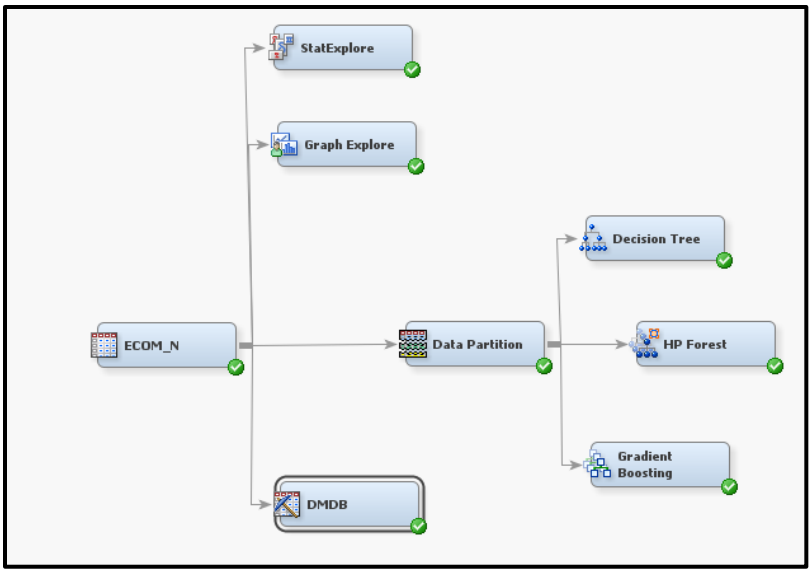
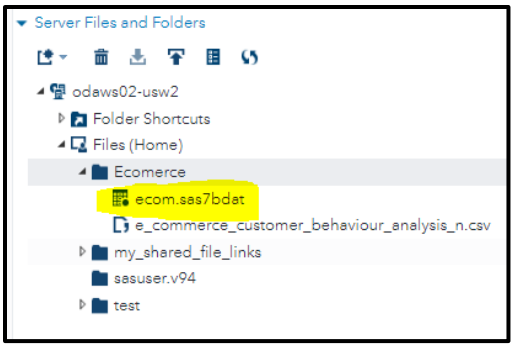


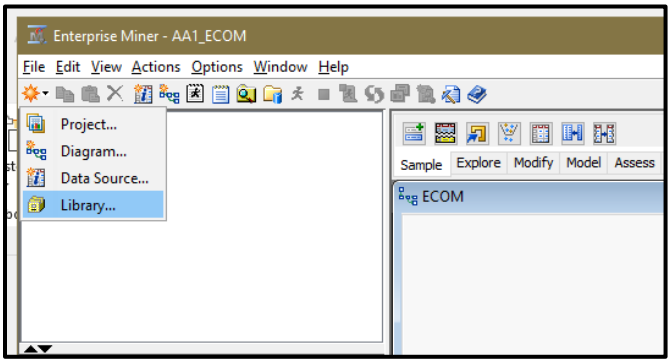

***Figure 1. Complete diagram of the flow***

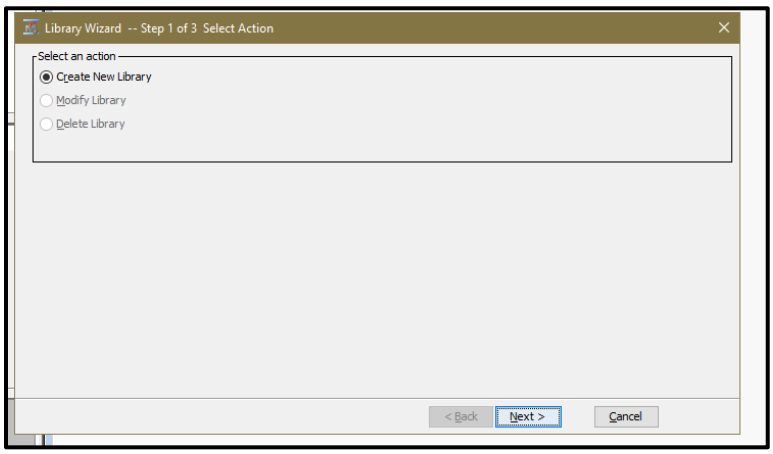
This flow diagram represents a process flow of a data mining project. Let's break down each component and its purpose:

1. **ECOM_N**: It representing the data source for the analysis.

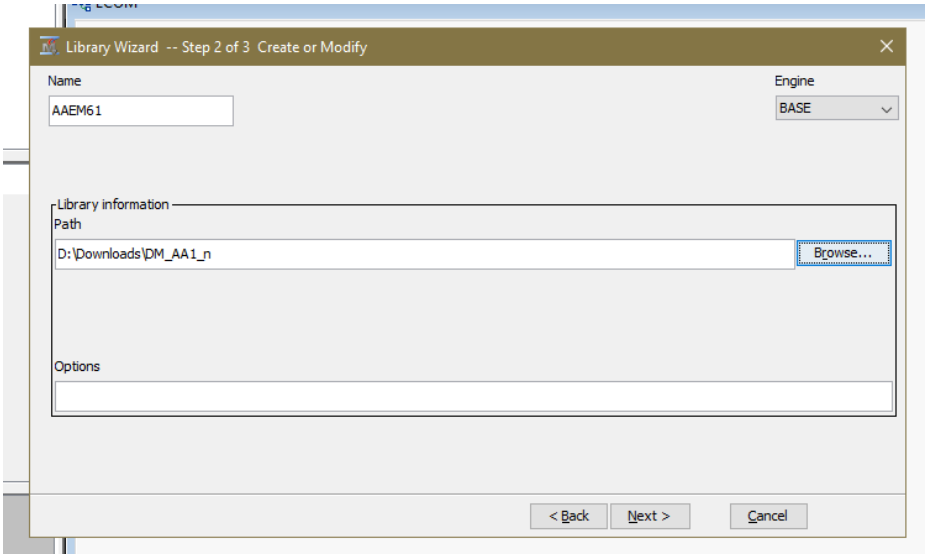


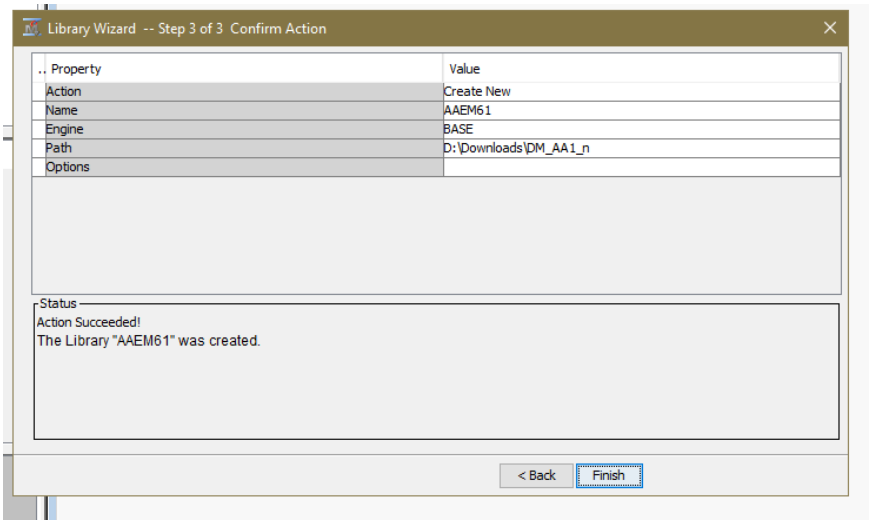

- The excel file was converted into a .sas7bdat format in order to be uploaded in the SAAS Enterprise Miner.
- To export the data we first have to create a library in SAAS Eterprise miner



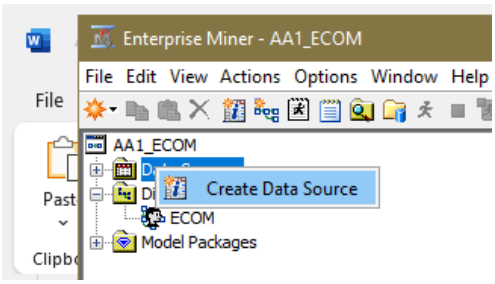



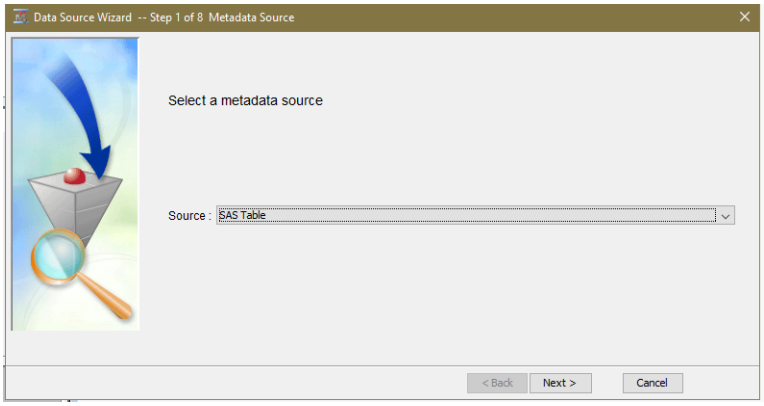
- We have to specify the name and path of the library to create.



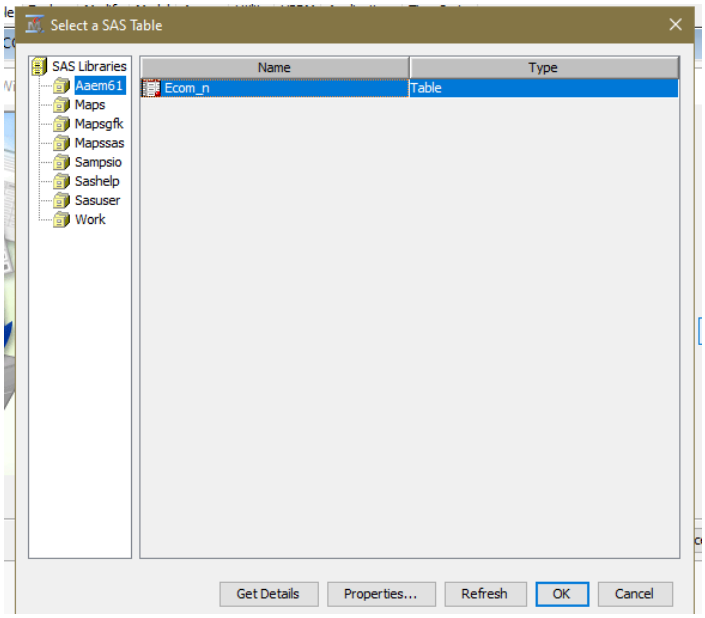



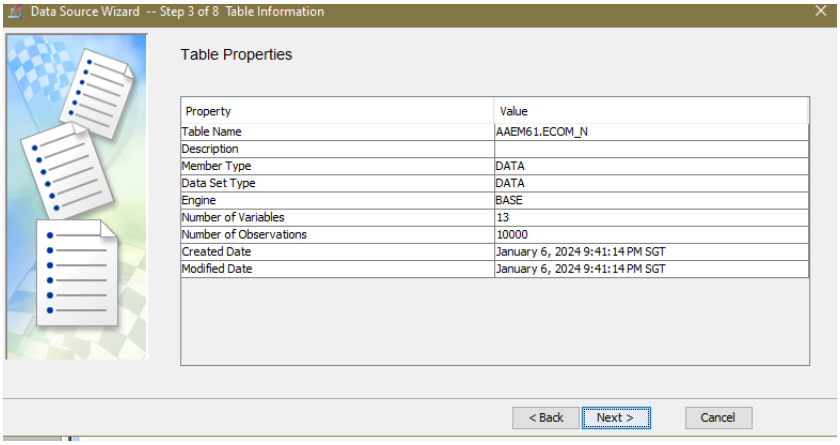

- After the library is created, we can now create the data source and import our data by clicking on create data source



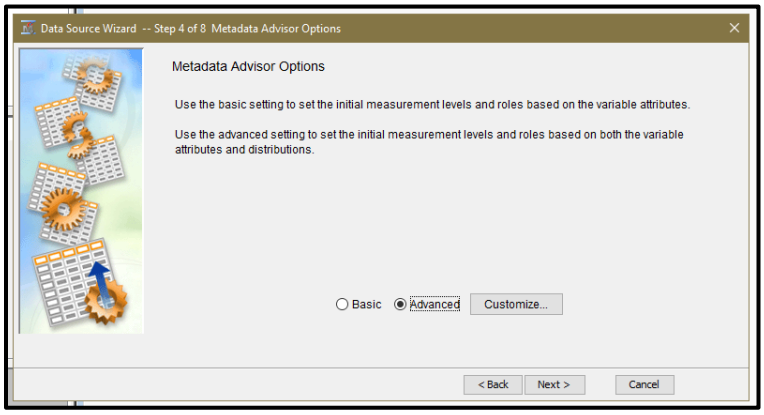



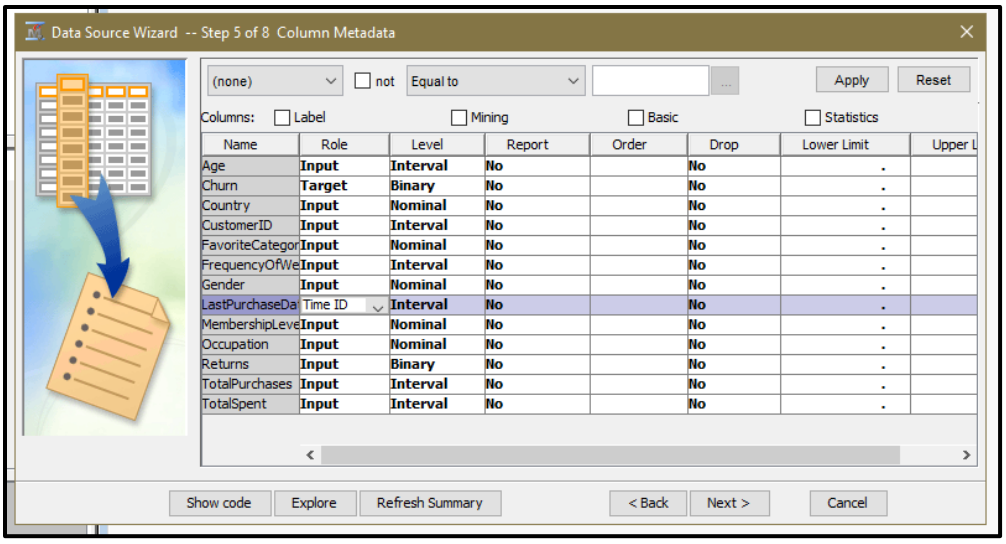

- Now have to choose the Ecom..sas7bdat file

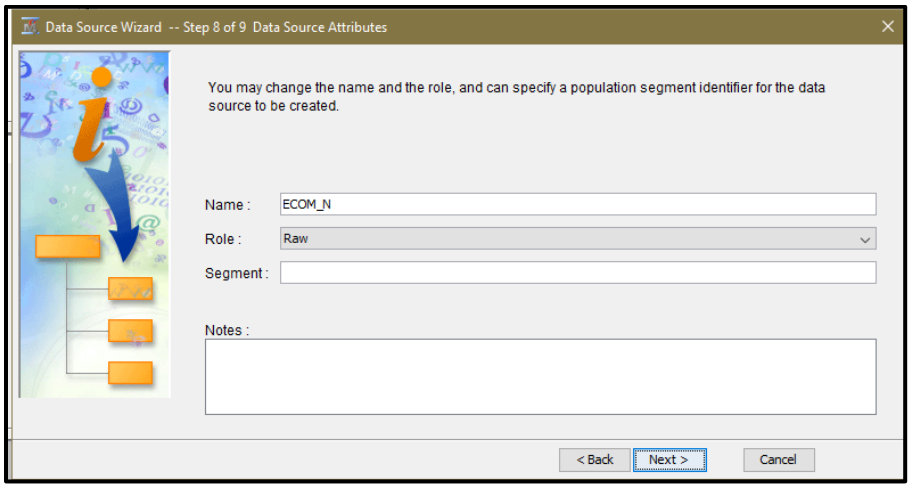




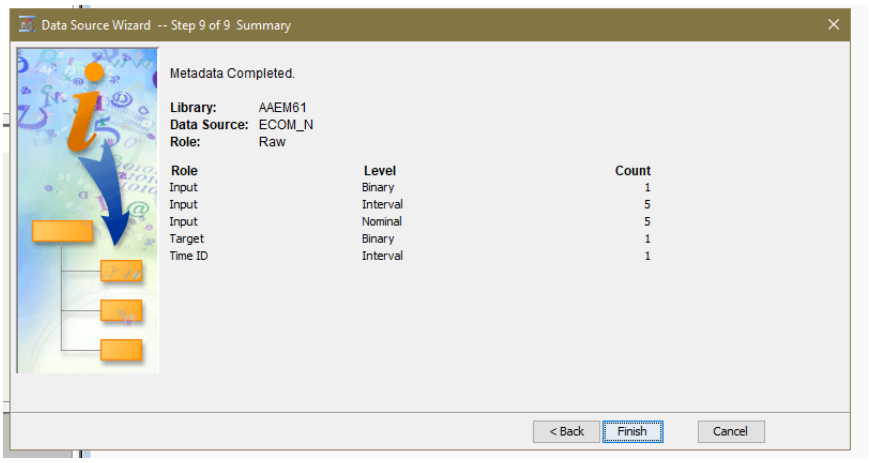

- By choosing the advance settings we can manually change the roles and level of the variables, as churn is our target variable, we have given it a target role and while all others have input role except Last LastPurchaseDate

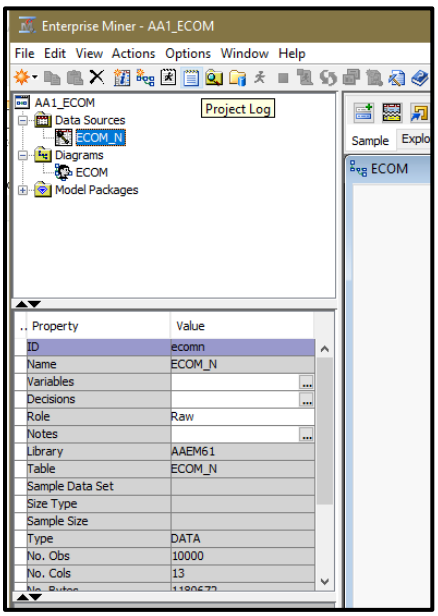





- The meta data is completed and the dataset is imported into SAS Enterprise Miner

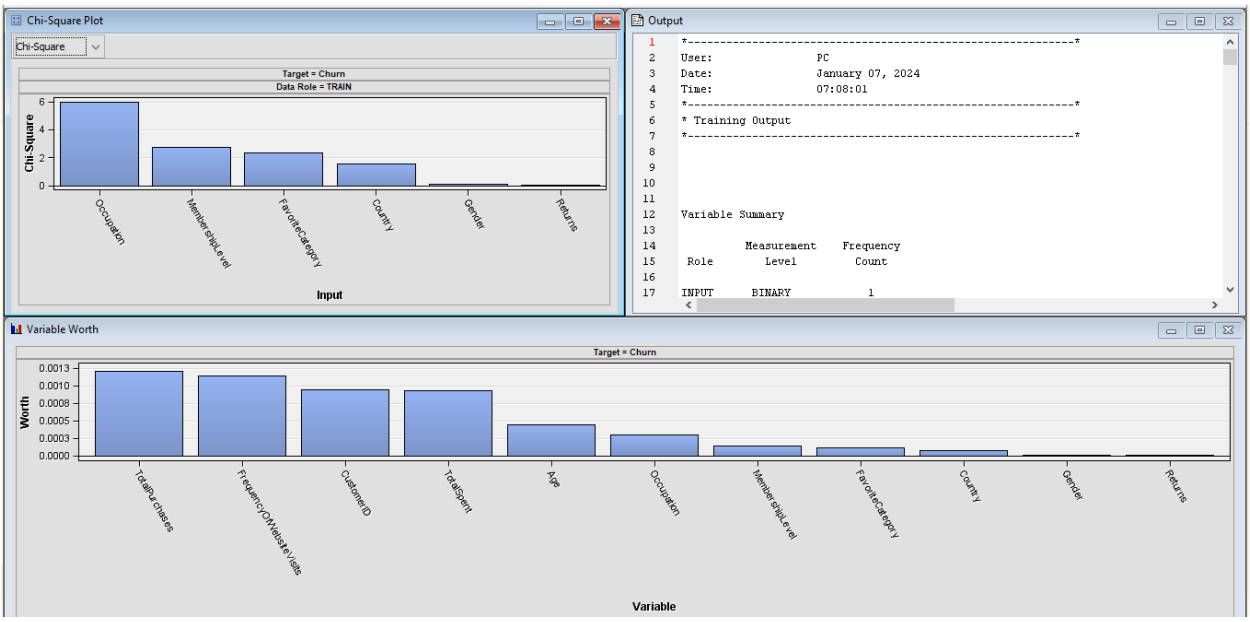




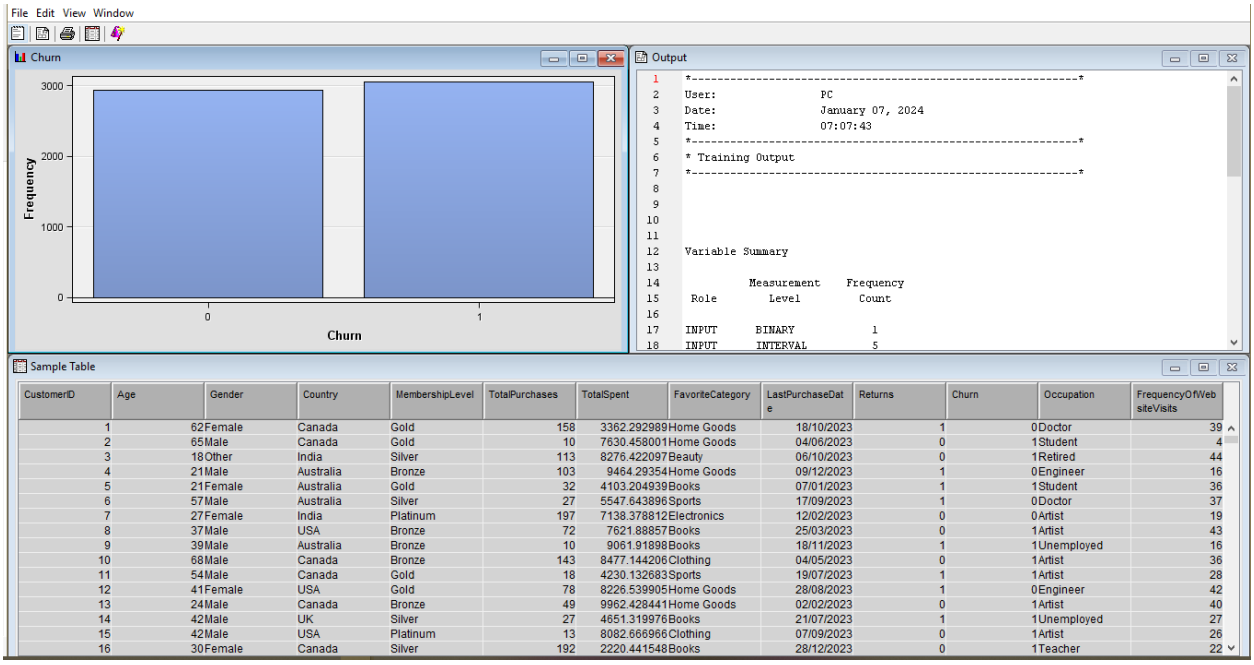

Class Variable Summary Statistics

| Variable | Label | Type | Number of Levels | Missing |
|---|---|---|---|---|
| Churn | | N | 2 | 0 |
| Country | | C | 5 | 0 |
| FavoriteCategory | | C | 6 | 0 |
| Gender | | C | 3 | 0 |
| MembershipLevel | | C | 4 | 0 |
| Occupation | | C | 7 | 0 |
| Returns | | N | 2 | 0 |

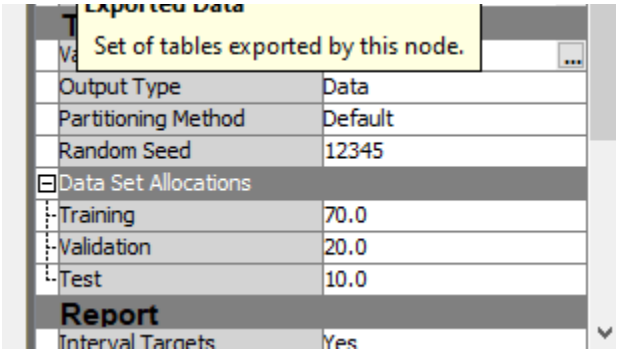
2. **StatExplore:** This node is used for exploratory statistical analysis. It provides summary statistics, frequency counts, and graphical analyses to understand the distribution and characteristics of the data.




3. **Graph Explore:** This node is for exploratory graphical analysis. It can be used to create various plots and charts that help in understanding the relationships between different variables in the dataset.



4. **Data Partition:** This is a crucial step in data mining where the dataset is divided into separate sets for training, validation, and testing. This allows for the evaluation of the model's performance and helps prevent overfitting.




- By clicking on the data partition node we can see the general properties panel.
- A training size of 70% was select from the side panel of general properties
- A Validation size of 20% was select from the side panel of general properties
- A Test size of 10 was select from the side panel of general properties

5. **Decision Tree:** This node represents a decision tree model. Decision trees are used for classification and regression tasks. They are easy to interpret and can handle both numerical and categorical data.

- The initial step is to connect the Imported data source to the Data partition node, it will partition the data into train, validation and test for the model input

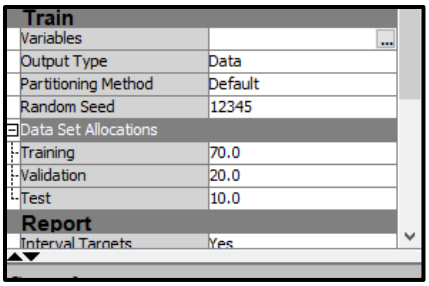
- Second step is to connect the data partition node to the decision tree node


- We can also change values of "max_depth" and "max_branch" of the decision tree algorithms from the general properties of the node, the parameters "max_depth" and "max_branch" are used to control the complexity and size of the tree, which, in turn, can affect the performance and generalization ability of the model in our case it reduced Misclassification rate from 0.6 to 0.4.9, the data partition is set to 70% training, 20 % validation and 10% for testing purpose.

Fit Statistics

| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| Churn | | _NOBS_ | Sum of Frequencies | 6998 | 2000 | 1002 |
| Churn | | _MISC_ | Misclassification Rate | 0.440126 | 0.466 | 0.499002 |
| Churn | | _MAX_ | Maximum Absolute Error | 0.916667 | 1 | 1 |
| Churn | | _SSE_ | Sum of Squared Errors | 3364.127 | 1024.491 | 519.8217 |
| Churn | | _ASE_ | Average Squared Error | 0.240363 | 0.256123 | 0.259392 |
| Churn | | _RASE_ | Root Average Squared E... | 0.490269 | 0.506086 | 0.509305 |
| Churn | | _DIV_ | Divisor for ASE | 13996 | 4000 | 2004 |
| Churn | | _DFT_ | Total Degrees of Freedom | 6998 | . | . |

| Property | Value |
|---|---|
| Splitting Rule | |
| Interval Target Criterion | Variance |
| Nominal Target Criterion | Gini |
| Ordinal Target Criterion | Gini |
| Significance Level | 0.2 |
| Missing Values | Use in search |
| Use Input Once | No |
| Maximum Branch | 3 |
| Maximum Depth | 5 |
| Minimum Categorical Size | 2 |
| Node | |
| Leaf Size | 8 |
| Number of Rules | 5 |
| Number of Surrogate Rules | 0 |
| Split Size | . |
| Split Search | |




6. **HP Forest:** HP stands for High-Performance. This node represents a random forest model, which is an ensemble learning method. It builds multiple decision trees and merges them together to get a more accurate and stable prediction.

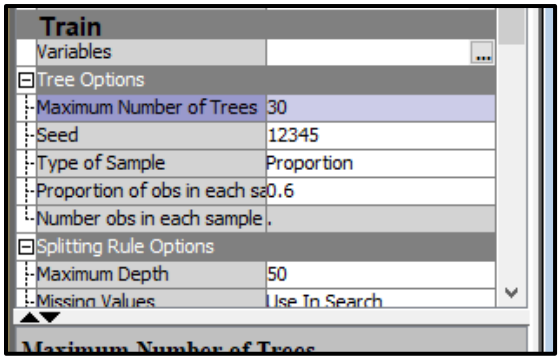
- The initial step is to drag the HP Forest component from the HPDM tab and then have to connect it with the data partition node. The data partition is set to 70% training, 20 % validation and 10% for testing purpose.


- Changing the HP forest properties can also help in increase results and performance. The “maximum number of trees” in the HP random forest model in SAS helps in achieving better results by providing a balance between model complexity and performance. By specifying the maximum number of trees, the model can capture complex patterns in the data, leading to improved predictive accuracy. However, it is essential to consider computational efficiency and the risk of overfitting. So that’s why

for our model we choose maximum number of trees as 30. It increased the accuracy and performance of the model.

Fit Statistics

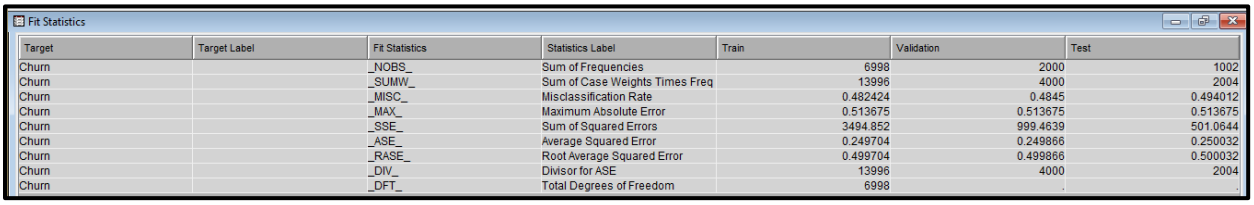
| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| Churn | | _ASE_ | Average Sq... | 0.248975 | 0.250009 | 0.25135 |
| Churn | | _DIV_ | Divisor for A... | 13996 | 4000 | 2004 |
| Churn | | _MAX_ | Maximum A... | 0.579557 | 0.579557 | 0.578166 |
| Churn | | _NOBS_ | Sum of Fre... | 6998 | 2000 | 1002 |
| Churn | | _RASE_ | Root Avera... | 0.498974 | 0.500009 | 0.501348 |
| Churn | | _SSE_ | Sum of Squ... | 3484.651 | 1000.035 | 503.7051 |
| Churn | | _DISF_ | Frequency ... | 6998 | 2000 | 1002 |
| Churn | | _MISC_ | Misclassific... | 0.466133 | 0.4975 | 0.515968 |
| Churn | | _WRONG_ | Number of ... | 3262 | 995 | 517 |




7. **Gradient Boosting:** This is another ensemble learning technique. Unlike random forests, gradient boosting builds one tree at a time, where each new tree helps to correct errors made by previously trained trees. It's often used for its predictive accuracy.

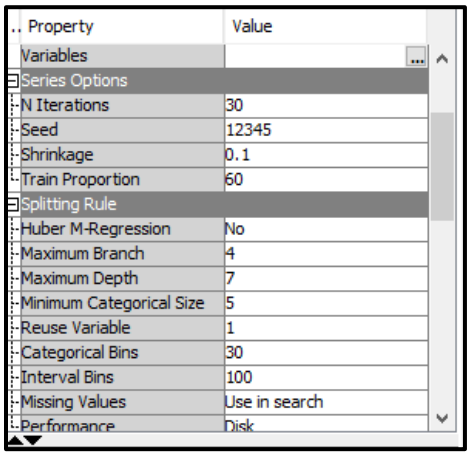
- SAS Enterprise Miner, the Gradient Boosting node is available on the Model tab of the toolbar, we can drag and connect it to the data partition node and
- The gradient boosting node provides various hyperparameters that can be tuned to optimize the model's performance.

- The maximum number of iterations, maximum branch, and maximum depth are hyperparameters that can help in achieving better results in Gradient Boosting in SAS Enterprise Miner. The maximum number of iterations determines the number of trees in the model, and increasing it can improve the model's performance. For our use case we kept the iterations as 30, cause if too much then if was causing misclassification too much.



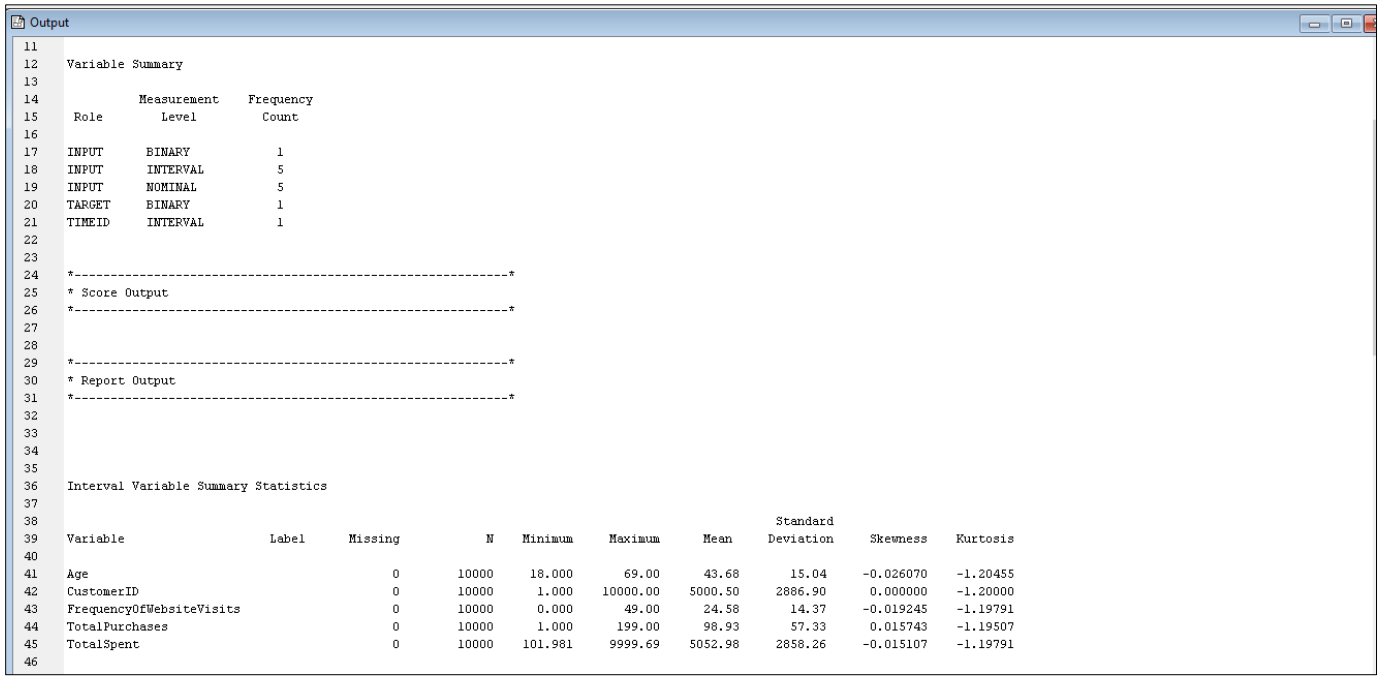
Fit Statistics

| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| Churn | | _NOBS_ | Sum of Frequencies | 6998 | 2000 | 1002 |
| Churn | | _SUMW_ | Sum of Case Weights Times Freq | 13996 | 4000 | 2004 |
| Churn | | _MISC_ | Misclassification Rate | 0.482424 | 0.4845 | 0.494012 |
| Churn | | _MAX_ | Maximum Absolute Error | 0.513675 | 0.513675 | 0.513675 |
| Churn | | _SSE_ | Sum of Squared Errors | 3494.852 | 999.4639 | 501.0644 |
| Churn | | _ASE_ | Average Squared Error | 0.249704 | 0.249866 | 0.250032 |
| Churn | | _RASE_ | Root Average Squared Error | 0.499704 | 0.499866 | 0.500032 |
| Churn | | _DIV_ | Divisor for ASE | 13996 | 4000 | 2004 |
| Churn | | _DFT_ | Total Degrees of Freedom | 6998 | . | . |



| Property | Value |
|---|---|
| Variables | |
| **Series Options** | |
| N Iterations | 30 |
| Seed | 12345 |
| Shrinkage | 0.1 |
| Train Proportion | 60 |
| **Splitting Rule** | |
| Huber M-Regression | No |
| Maximum Branch | 4 |
| Maximum Depth | 7 |
| Minimum Categorical Size | 5 |
| Reuse Variable | 1 |
| Categorical Bins | 30 |
| Interval Bins | 100 |
| Missing Values | Use in search |
| Performance | Disk |

8. **DMDb**: This could be a database node or a data mining database reference. It usually points to a dataset that's been preprocessed and is ready for mining or has been used to store processed data.



Output

Variable Summary

| Role | Measurement Level | Frequency Count |
|---|---|---|
| INPUT | BINARY | 1 |
| INPUT | INTERVAL | 5 |
| INPUT | NOMINAL | 5 |
| TARGET | BINARY | 1 |
| TIMEID | INTERVAL | 1 |

*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*

*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*

Interval Variable Summary Statistics

| Variable | Label | Missing | N | Minimum | Maximum | Mean | Standard Deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| Age | | 0 | 10000 | 18.000 | 69.00 | 43.68 | 15.04 | -0.026070 | -1.20455 |
| CustomerID | | 0 | 10000 | 1.000 | 10000.00 | 5000.50 | 2886.90 | 0.000000 | -1.20000 |
| FrequencyOfWebsiteVisits | | 0 | 10000 | 0.000 | 49.00 | 24.58 | 14.37 | -0.019245 | -1.19791 |
| TotalPurchases | | 0 | 10000 | 1.000 | 199.00 | 98.93 | 57.33 | 0.015743 | -1.19507 |
| TotalSpent | | 0 | 10000 | 101.981 | 9999.69 | 5052.98 | 2858.26 | -0.015107 | -1.19791 |