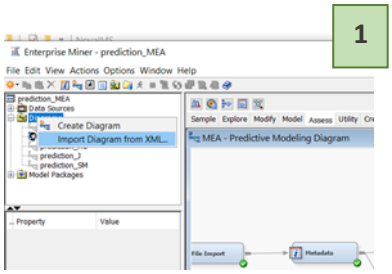
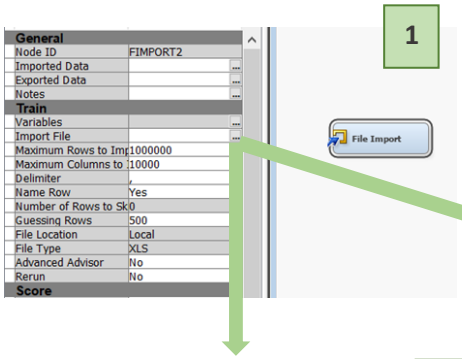
12 - Diagram Importation

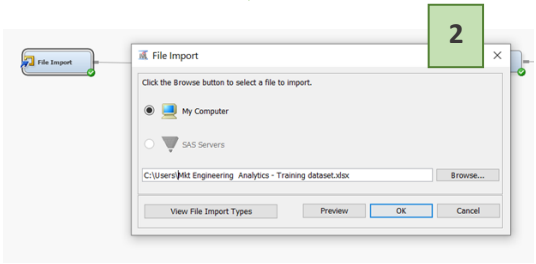


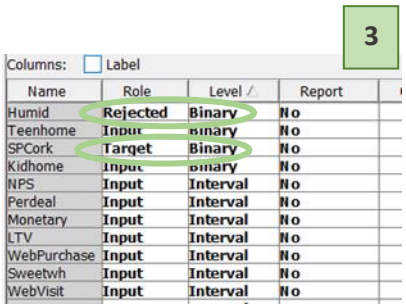
To import a diagram from an XML file you just need to right click on Diagrams option.

13 – File Import (to import training data set file and to define the target variable)

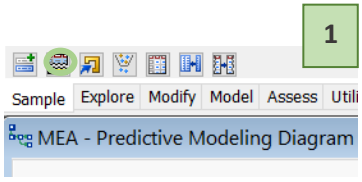


After the file importation, it is necessary to define the role and level for the target variable. Also, reject the one that you do not want to include on your analysis.

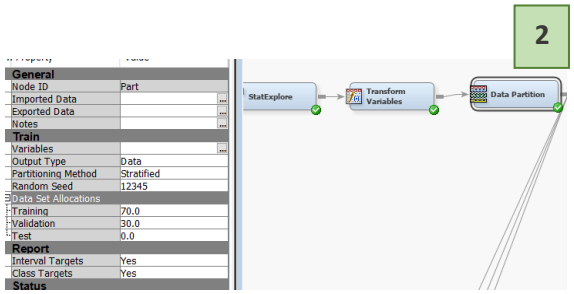




14 – Data Partition



Drag and drop the Data Partition node from Sample tab and define the percentage of individuals that go for each of the sets defined on the properties panel.

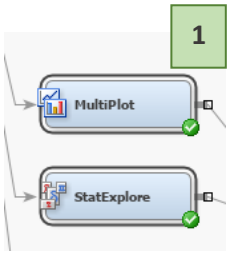


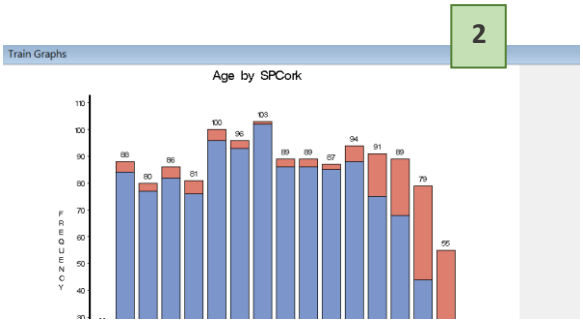
Partitioning Method - to specify the sampling method to use in order to partition the data:

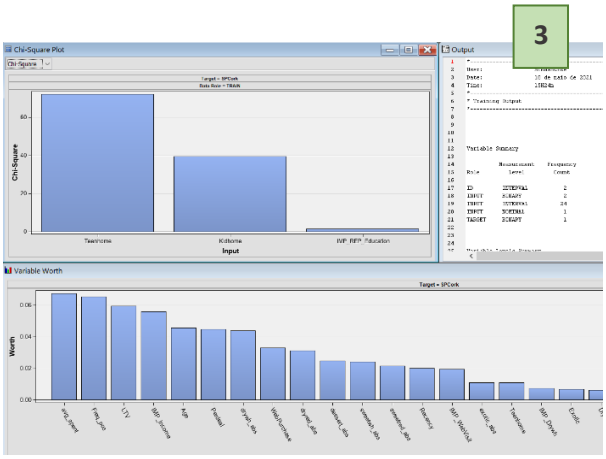
Stratified - specify variables to form strata (or subgroups) of the total population. Within each stratum, all observations have an equal probability of being written to one of the partitioned data sets.

15 – Descriptive Analysis (Multiplot + Stat Explore)

Drag and drop a Multiplot and Stat Explore node to explore your data regarding the inclusion of the target variable. Link it to the Data Partition node.







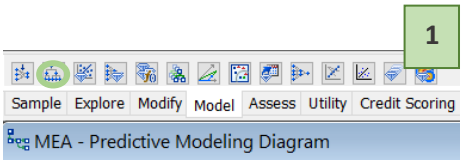
The **Train Graph** window allows you to do a comparison between customers that bought the product and those who did not, i.e., you can trace a profile for both types of customers.

Red colour represents the ones that bought the product.

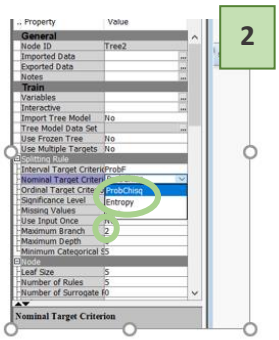
Observing the **Chi-Square Plot** it is possible to know which categorical variable explains better the data.

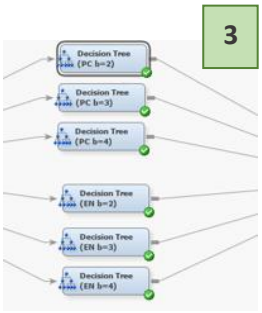
Using **Variable Worth** output you can observe which variables have a higher discriminant power in explaining the target variable. This will be useful to decide which variables you want the program to consider when modelling regression and neural networks.

16 – Decision Tree



Drag and drop the Decision Tree node from the Model tab. Link it to the Data Partition node.





To define different types of decision trees, you have to manipulate the properties panel options regarding the Nominal Target Criterion and the number of Maximum Branches.

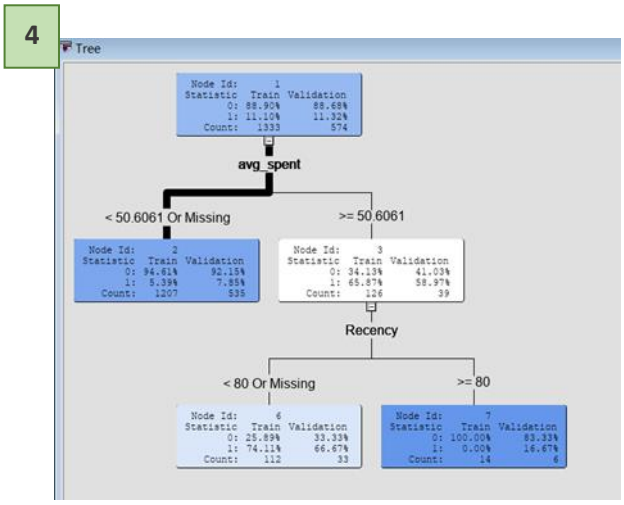
Nominal Target Criterion - here you define which method you want to apply to the tree. These methods define the purity of the tree and which attributes to select.

ProbChisq — the p-value of the Pearson Chi-square statistic for the target versus the branch node;

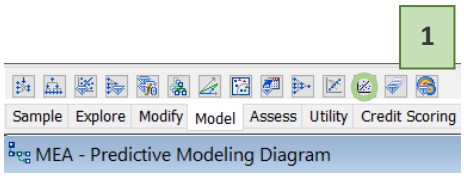
Entropy — reduction in the entropy measure.

Maximum Branch - defines how many branches we want for the tree. Also, this defines the complexity of the tree. If you define a high number of branches some of them could define specificities of the training data, contributing for overfitting.

After running, the results' window allows you to observe the tree created.

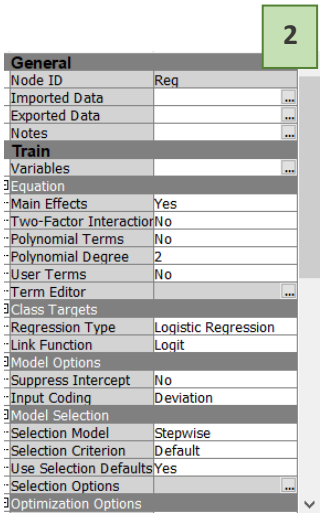


17 – Regression



Drag and drop the Regression node from the Model tab.

Before this step, do not forget to include a Metadata node to choose the variables that you want the program to consider for Regression and Neural Networks. You want only the variables that best describe the target.



To define regression, in this case, you have to define:

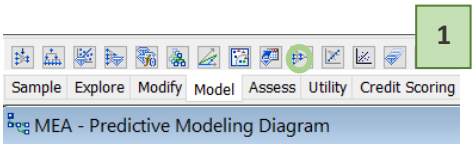
Regression Type - to specify the type of regression that you want:

Logistic Regression - the default regression type for binary or ordinal targets.

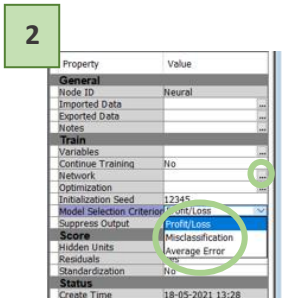
Link Function - to specify the link function. Link functions link the response mean to the linear predictor.

Logit - (default) Specifies the inverse of the cumulative logistic distribution function.

18 – Neural Network



Drag and drop the Neural Network node from the Model tab. Link it to a Metadata node.



To define different types of neural networks, you have to manipulate the properties panel options regarding the Model Selection Criterion.

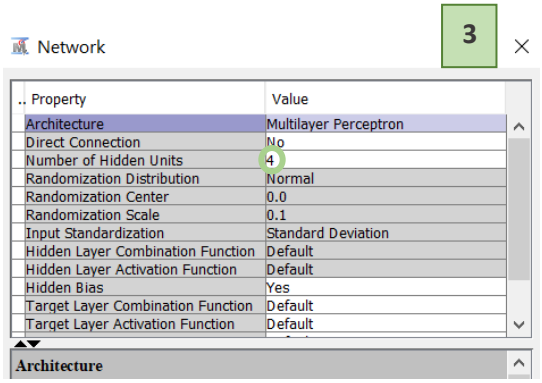
Model Selection Criterion - to specify the most appropriate model from your network:

Profit/loss - chooses the model that maximizes the profit or minimizes the loss;

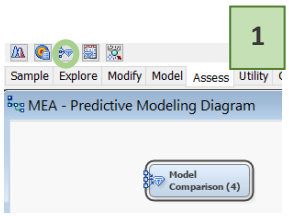
Misclassification - the model that has the smallest misclassification rate;

Average error - chooses the model that has the smallest average error.

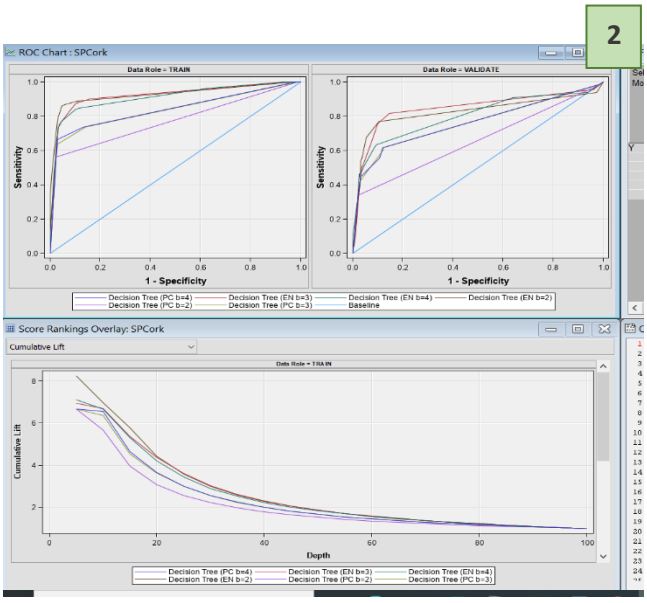
Also, clicking on the Network option a pop-up window appears and in there you can define the number of hidden layers to apply.



19 – Model Comparison (to compare models – helps to decide which one is the best one)



Drag and drop the Model Comparison node from the Assess tab. Link it to all the methods previously defined an run it.



In the results it is possible to explore some interesting outputs:

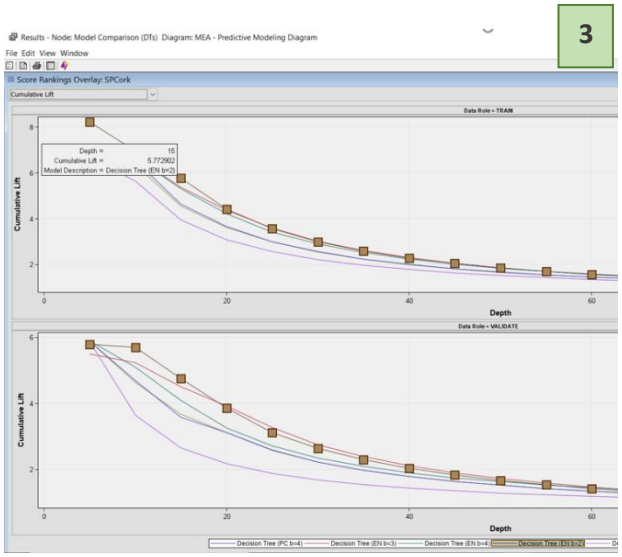
ROC Chart - displays multiple ROC curves* for competing models on a single chart.

Score Rankings - displays the score rankings plots for the competing models.

*The **ROC curve** shows the trade-off between sensitivity and specificity.

Expanding the Score Rankings window there are 3 types of graphics that gives you useful information: cumulative lift; cumulative % response; cumulative & captured response. Cumulative lift is the one that allows you to decide which model has the best performance.

3

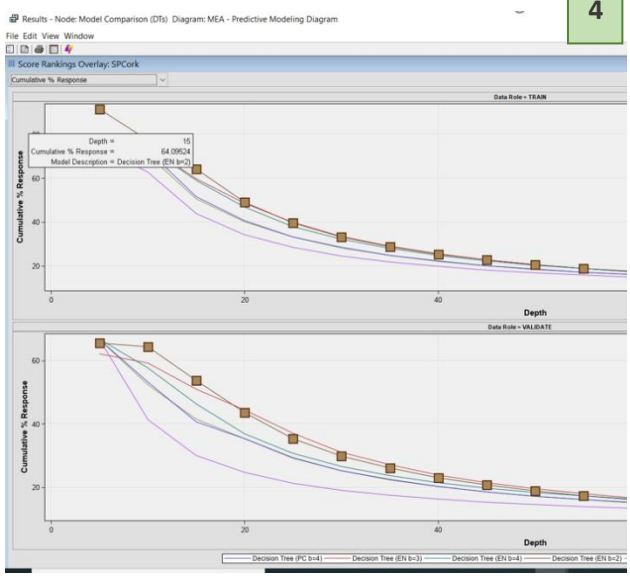


Cumulative Lift indicates how many times better is your model in comparison with a baseline.

The customers are order by their probability of buying the product. The ones with higher scores appears first.

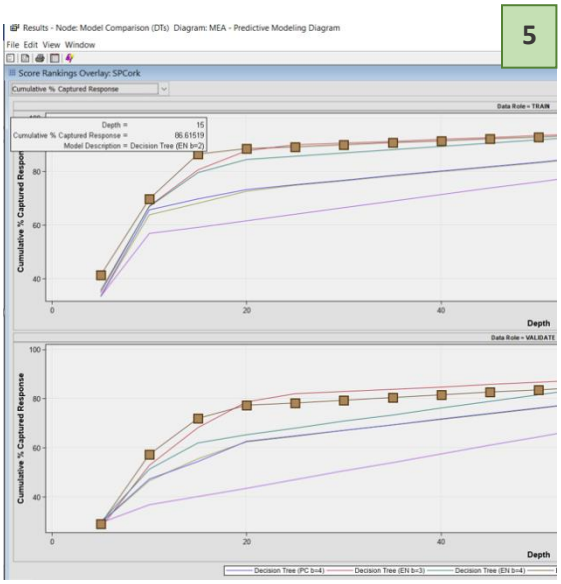
Example: If I choose to contact the 15% of the total customers that the model indicates me, my results will be 5.77 times better than contact randomly 15% of the total customers.

4



Cumulative % Response indicates the percentage of customers that buy the product.

Example: If I choose to contact the 15% of the total customers that the model indicates me, 64% of them buy the product.



Cumulative % Captured Response indicates the percentage of customers that buy the product that it is possible to capture, regarding all the customers that buy the product.

Example: If I choose to contact the 15% of the total customers that the model indicates me, I will capture 86.61% of the total customers that buy the product.

20 – Export data from a node