Neural Network with Claim Fraud Dataset/ SAS Enterprise Miner

Didem B. Aykurt

Colorado State University Global

MIS530; Predictive Analytics

Dr.Jennifer Catalano

February 26, 2023

Neural Network

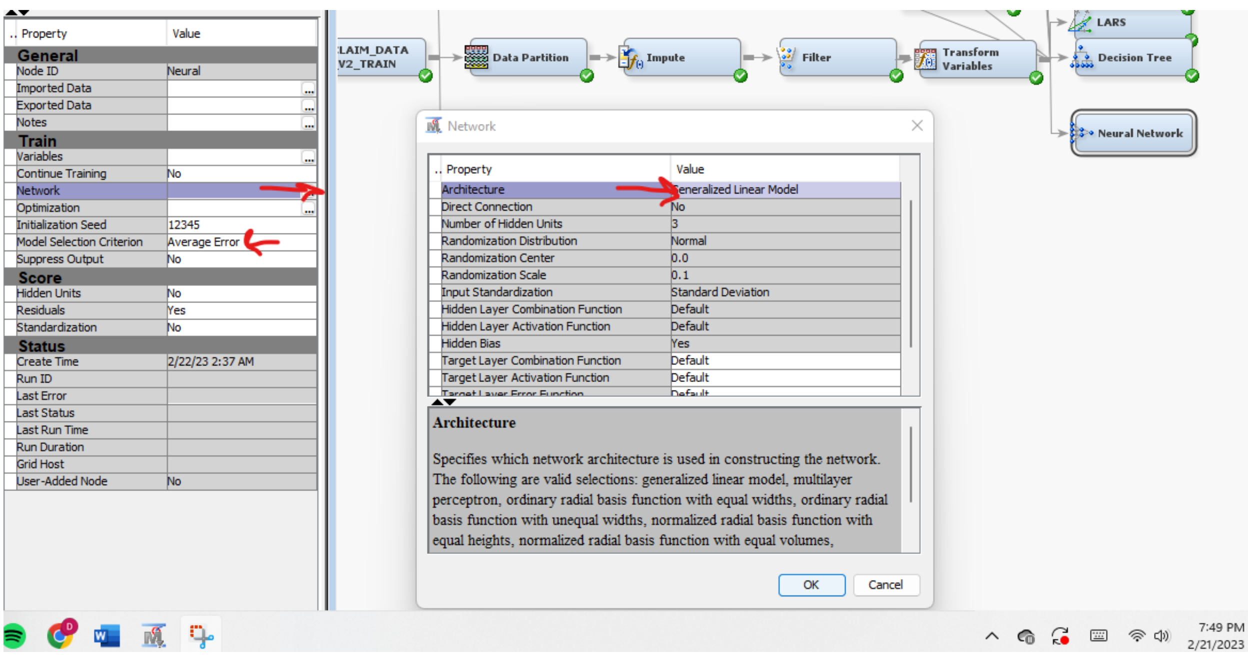
Neural networks are great models to analyze and mimic functions like the human brain as humans learn their experiences. The neural network model should also learn, like brain neurons build cognition and intelligence, how a neural network learns that the training dataset should be sufficiently large enough for the building network to calculate the value of each node that was the model loading all detail learn. The brain network of neurons works with one cohesive unit. Inputs come to each neuron through a dendrite connection, which helps send information to the neurotransmitters across a synaptic gap. The number of neurotransmitters that transfer information fast and substantially, so if the number of high synaptic gaps has strength relative to each dendrite's connection to their response. Additionally, if the synapse is more active, that means strong ties; otherwise, the weaker synapse lacks the use. "When an axon of cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."(Hebb,1949) That means the combination of neurons' actions together, which is a strong connection between two neurons, could be adjusted. If the adjustment is low, the result might be a very long training time. However, if it is high, the result might be a variance from the aspiration solution.

SAS Enterprise Miner has a Neural network node to handle datasets with multiple target variables. A class of target variables results in a probability and an interval target results in an expected value. The neural network model is an excellent benefit for a complex set of nonlinear models; it transforms the variables into a model estimation. SAS Enterprise Miner uses the formula by the hidden layers that help to know the hidden layer combination function. Also, the node has a function to specify a target layer combination function to show how the inputs might be combined.

Creating a Neural Network Node

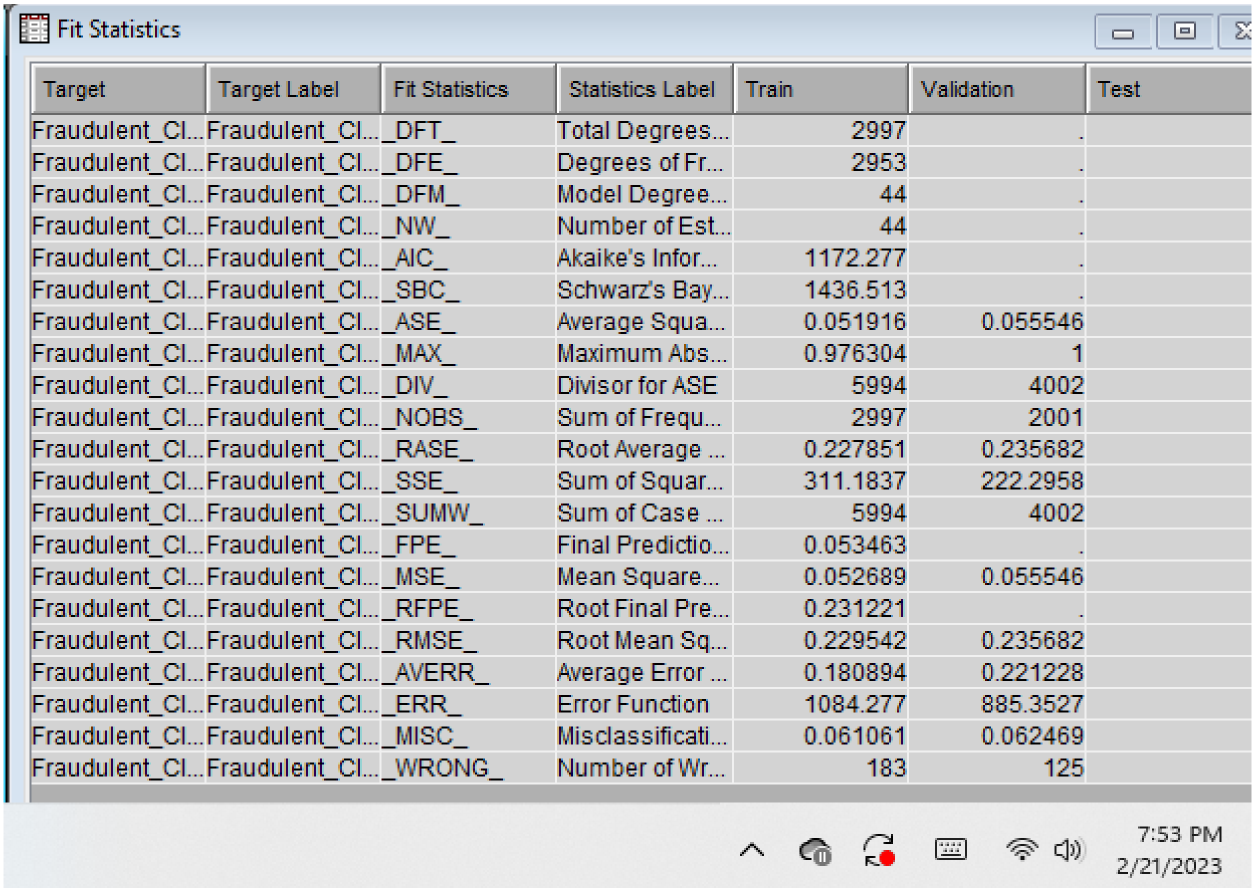
I will use an automobile insurance claim dataset to apply a few neural networks and compare the results to see which neural network is best to predict. The first one is the neural network generalized linear model. Drag and drop the neural network node in the Model tab as the model's name from its default on Node ID will Neural3. Click the Network ellipse to customize the network model. I will choose the Generalized Linear Model. Model Selection Criterion to Average Error. Target Layer combination, activation, and error set Default. And Run node.

Figure 1: Neural network node and property.



The result of the fit statistics window shows an average square error of 0.0555546, which results higher than the system-generated decision tree average square error of 0.053507. The decision tree is slightly better than a neural network- a general linear method.

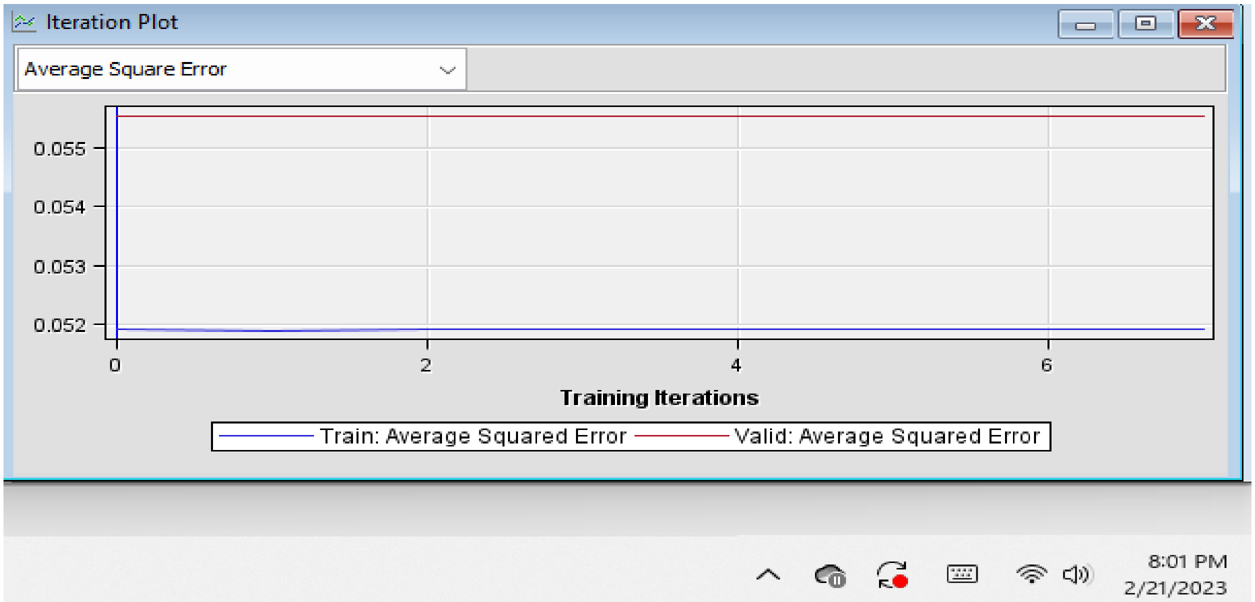
Figure 2: Fit Statistics for the generalized linear model neural network



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Cl...	Fraudulent_Cl...	_DFT_	Total Degrees...	2997	.	
Fraudulent_Cl...	Fraudulent_Cl...	_DFE_	Degrees of Fr...	2953	.	
Fraudulent_Cl...	Fraudulent_Cl...	_DFM_	Model Degree...	44	.	
Fraudulent_Cl...	Fraudulent_Cl...	_NW_	Number of Est...	44	.	
Fraudulent_Cl...	Fraudulent_Cl...	_AIC_	Akaike's Infor...	1172.277	.	
Fraudulent_Cl...	Fraudulent_Cl...	_SBC_	Schwarz's Bay...	1436.513	.	
Fraudulent_Cl...	Fraudulent_Cl...	_ASE_	Average Squa...	0.051916	0.055546	
Fraudulent_Cl...	Fraudulent_Cl...	_MAX_	Maximum Abs...	0.976304	1	
Fraudulent_Cl...	Fraudulent_Cl...	_DIV_	Divisor for ASE	5994	4002	
Fraudulent_Cl...	Fraudulent_Cl...	_NOBS_	Sum of Frequ...	2997	2001	
Fraudulent_Cl...	Fraudulent_Cl...	_RASE_	Root Average ...	0.227851	0.235682	
Fraudulent_Cl...	Fraudulent_Cl...	_SSE_	Sum of Squar...	311.1837	222.2958	
Fraudulent_Cl...	Fraudulent_Cl...	_SUMW_	Sum of Case ...	5994	4002	
Fraudulent_Cl...	Fraudulent_Cl...	_FPE_	Final Predictio...	0.053463	.	
Fraudulent_Cl...	Fraudulent_Cl...	_MSE_	Mean Square...	0.052689	0.055546	
Fraudulent_Cl...	Fraudulent_Cl...	_RFPE_	Root Final Pre...	0.231221	.	
Fraudulent_Cl...	Fraudulent_Cl...	_RMSE_	Root Mean Sq...	0.229542	0.235682	
Fraudulent_Cl...	Fraudulent_Cl...	_AVERR_	Average Error ...	0.180894	0.221228	
Fraudulent_Cl...	Fraudulent_Cl...	_ERR_	Error Function	1084.277	885.3527	
Fraudulent_Cl...	Fraudulent_Cl...	_MISC_	Misclassificati...	0.061061	0.062469	
Fraudulent_Cl...	Fraudulent_Cl...	_WRONG_	Number of Wr...	183	125	

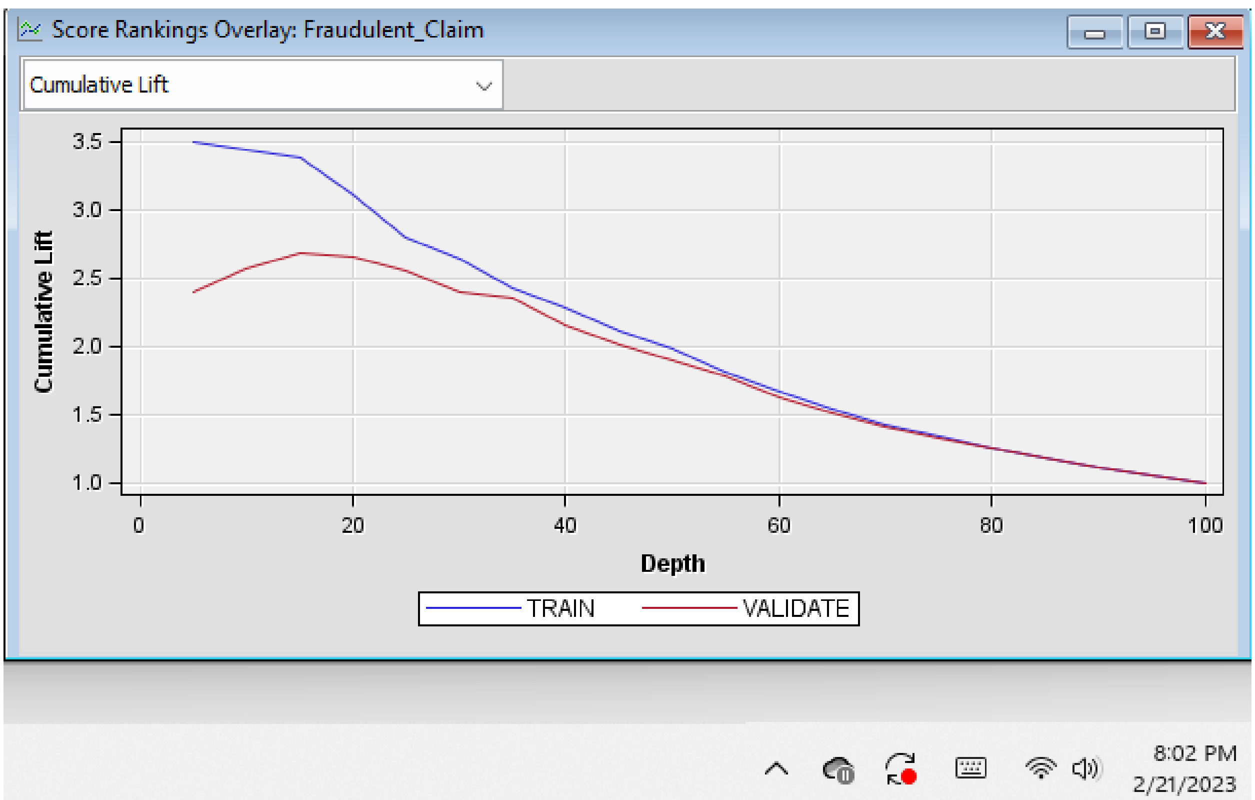
The Iteration Plot window explains how average square error changes training iteration. In this case, it has six iterations, and the model strengthens very quickly as the average square error shows no improvement in any iteration.

Figure 3: Iteration Plot for the generalized linear model neural network.



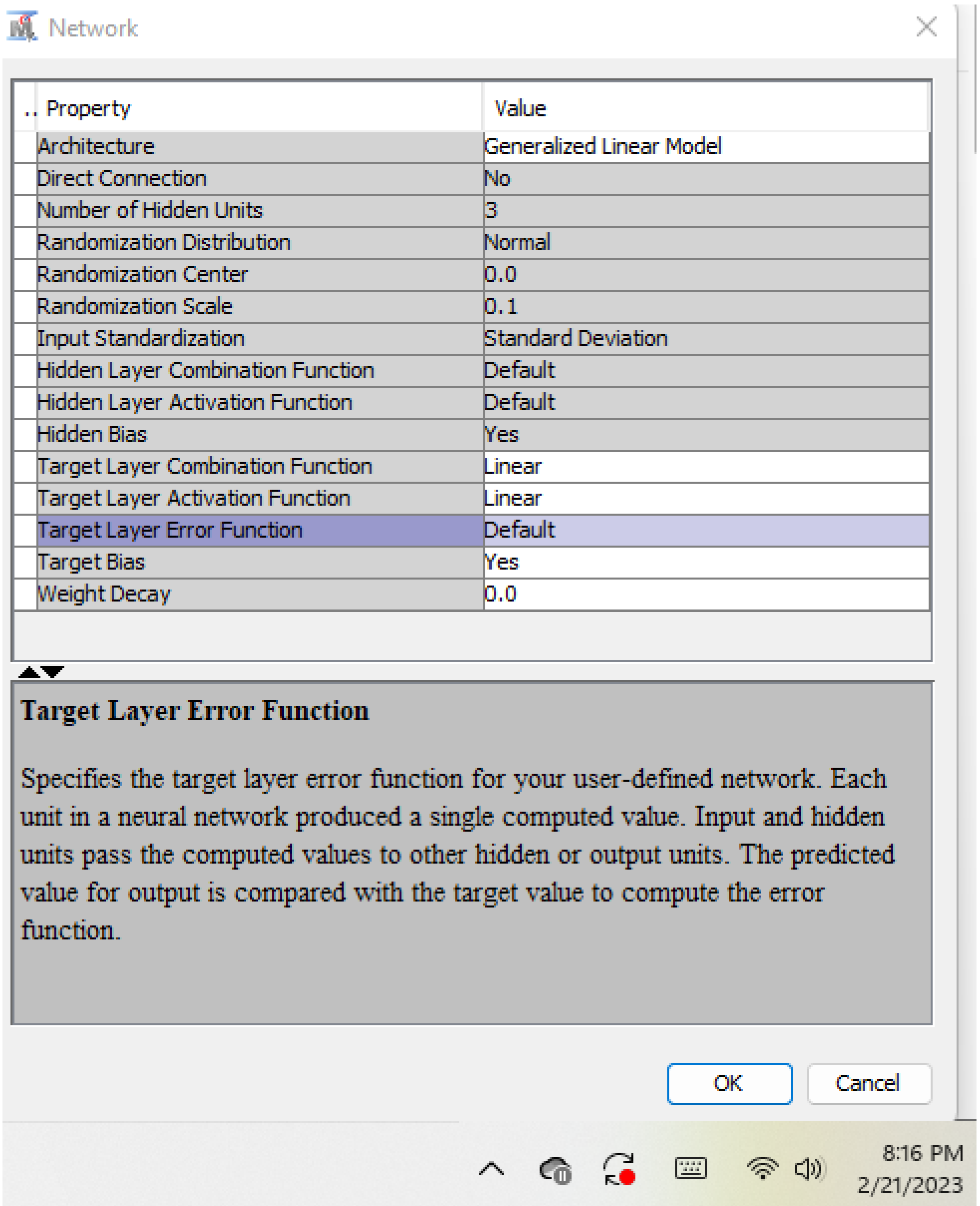
In the first step of 15% of data, the cumulative lift is over 3.38, which signals the strangeness of the GLM model.

Figure 4: Cumulative Lift for the generalized linear model neural network.



Let us look at what happens when the target activation and combination function are set to linear.

Figure 5: Network window for the generalized linear model neural network.



The result of an average square error at 0.059704 is higher than the target activation, and the combination was a default function, so this network has a worse effect. Therefore, I will not use it any further in this case.

Figure 5: Fit Statistics window for the generalized linear model neural network.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Cl...	Fraudulent_Cl...	_DFT_	Total Degrees...	2997	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_DFE_	Degrees of Fr...	2909	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_DFM_	Model Degree...	88	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_NW_	Number of Est...	88	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_AIC_	Akaike's Infor...	1128.424	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_SBC_	Schwarz's Bay...	1656.896	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_ASE_	Average Squa...	0.058845	0.059704	.
Fraudulent_Cl...	Fraudulent_Cl...	_MAX_	Maximum Abs...	0.99685	0.996979	.
Fraudulent_Cl...	Fraudulent_Cl...	_DIV_	Divisor for ASE	5994	4002	.
Fraudulent_Cl...	Fraudulent_Cl...	_NOBS_	Sum of Frequ...	2997	2001	.
Fraudulent_Cl...	Fraudulent_Cl...	_RASE_	Root Average ...	0.242579	0.244344	.
Fraudulent_Cl...	Fraudulent_Cl...	_SSE_	Sum of Squar...	352.7153	238.9357	.
Fraudulent_Cl...	Fraudulent_Cl...	_SUMW_	Sum of Case ...	5994	4002	.
Fraudulent_Cl...	Fraudulent_Cl...	_FPE_	Final Predictio...	0.062405	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_MSE_	Mean Square...	0.060625	0.059704	.
Fraudulent_Cl...	Fraudulent_Cl...	_RFPE_	Root Final Pre...	0.24981	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_RMSE_	Root Mean Sq...	0.246221	0.244344	.
Fraudulent_Cl...	Fraudulent_Cl...	_AVERR_	Average Error ...	0.158896	20.55991	.
Fraudulent_Cl...	Fraudulent_Cl...	_ERR_	Error Function	952.4239	82280.76	.
Fraudulent_Cl...	Fraudulent_Cl...	_MISC_	Misclassificati...	0.061061	0.061969	.
Fraudulent_Cl...	Fraudulent_Cl...	_WRONG_	Number of Wr...	183	124	.

8:18 PM
2/21/2023

Let us compare Multilayer Perceptron and GLM neural networks. First, set the Multilayer Perceptron function on Architecture and the default function for target layer activation and combination.

Figure 6: Network property window for the Multilayer Perception model neural network.

Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	3
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default
Target Bias	Yes
Weight Decay	0.0

Architecture

Specifies which network architecture is used in constructing the network. The following are valid selections: generalized linear model, multilayer perceptron, ordinary radial basis function with equal widths, ordinary radial basis function with unequal widths, normalized radial basis function with equal heights, normalized radial basis function with equal volumes, normalized radial basis function with equal widths, normalized radial basis function with equal widths and heights, normalized radial basis function with unequal widths and heights and a User specified network.

12:04 PM
2/22/2023

The average square error of 0.056747 is higher than GLM's result of an average square error of 0.0555546. The GLM is slightly better than MLP for the claim fraud dataset.

Figure 7: Fit Statistics window for the Multilayer Perception model neural network.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudule...	Fraudule...	DFT	Total Degrees ...	2997	.	
Fraudule...	Fraudule...	DFE	Degrees of Fre...	2861	.	
Fraudule...	Fraudule...	DFM	Model Degrees...	136	.	
Fraudule...	Fraudule...	NW	Number of Esti...	136	.	
Fraudule...	Fraudule...	AIC	Akaike's Inform...	1293.691	.	
Fraudule...	Fraudule...	SBC	Schwarz's Bay...	2110.421	.	
Fraudule...	Fraudule...	ASE	Average Squar...	0.050137	0.056747	
Fraudule...	Fraudule...	MAX	Maximum Abso...	0.961283	0.999899	
Fraudule...	Fraudule...	DIV	Divisor for ASE	5994	4002	
Fraudule...	Fraudule...	NOBS	Sum of Freque...	2997	2001	
Fraudule...	Fraudule...	RASE	Root Average ...	0.223912	0.238216	
Fraudule...	Fraudule...	SSE	Sum of Square...	300.5189	227.1015	
Fraudule...	Fraudule...	SUMW	Sum of Case ...	5994	4002	
Fraudule...	Fraudule...	FPE	Final Prediction...	0.054903	.	
Fraudule...	Fraudule...	MSE	Mean Squared ...	0.05252	0.056747	
Fraudule...	Fraudule...	RFPE	Root Final Pre...	0.234314	.	
Fraudule...	Fraudule...	RMSE	Root Mean Squ...	0.229172	0.238216	
Fraudule...	Fraudule...	AVERR	Average Error ...	0.170452	0.203379	
Fraudule...	Fraudule...	ERR	Error Function	1021.691	813.9239	
Fraudule...	Fraudule...	MISC	Misclassificatio...	0.059726	0.067466	
Fraudule...	Fraudule...	WRON...	Number of Wro...	179	135	

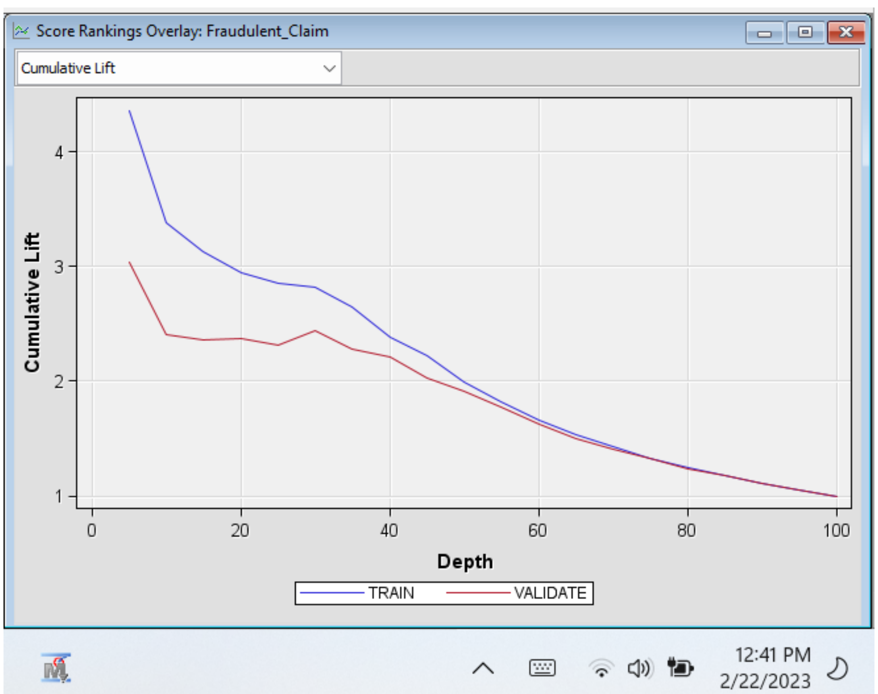
12:09 PM
2/22/2023

There are 50 training iterations as the model slowly iterated as the average square error shows an increase until 50, so the model is not statistically significant.

Figure 8: Iteration Plot window for the Multilayer Perception model neural network.

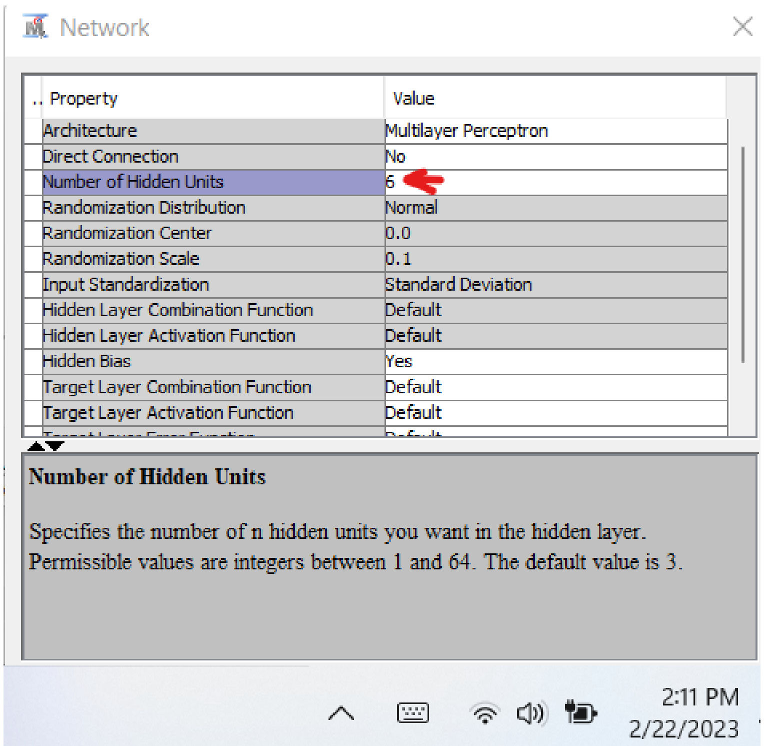
In the first step of 15% of data, the cumulative lift is over 3.12, which signals higher than GLM.

Figure 9: Cumulative Lift window for the Multilayer Perception model neural network.



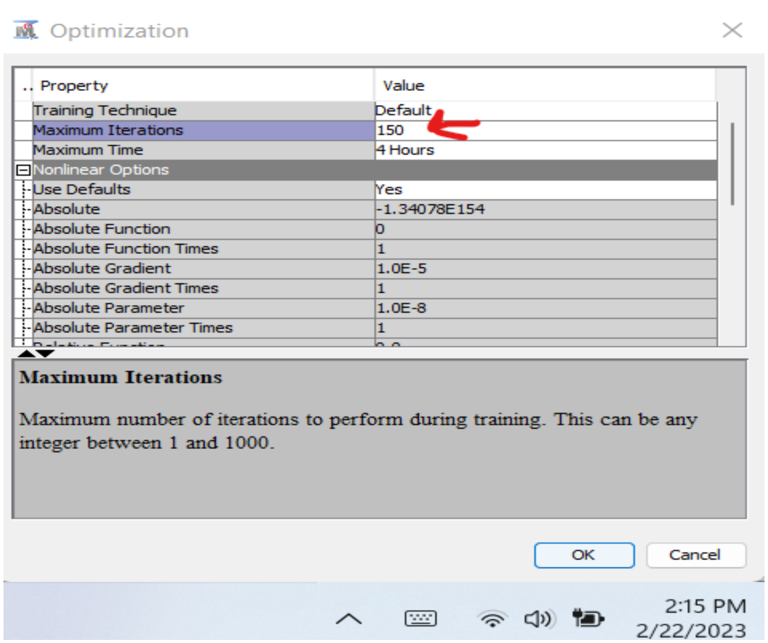
The neural network will use the MLP function on Architecture, and the number of hidden units will increase by 3 to 6, which means increasing the complexity of the network.

Figure 10: Network window set several hidden units six for the Multilayer Perception model neural network.



Click ellipse on Optimization, then pop up a new window to set maximum iteration 50 to 150, increasing the number of relations between the nodes.

Figure 11: Optimization window to set iteration number 150 for the Multilayer Perception model neural network.



Adjusting the neural network's complexity affected the result of an average square error of 0.055845; this is slightly lower than the three remote unit networks by PLM average square error of 0.056747.

Figure 12: Fit Statistics sets the number of hidden units to six and the iteration number to 150 for the multilayer perception model neural network.

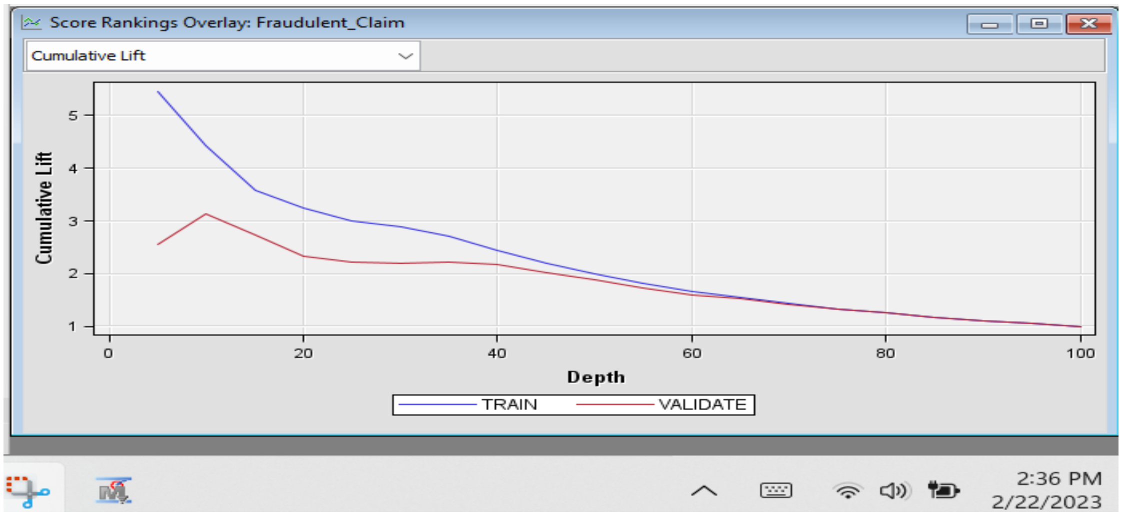
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudule...	Fraudule...	DFT	Total De...	2997	.	
Fraudule...	Fraudule...	DFE	Degrees ...	2726	.	
Fraudule...	Fraudule...	DFM	Model De...	271	.	
Fraudule...	Fraudule...	NW	Number ...	271	.	
Fraudule...	Fraudule...	AIC	Akaike's I...	1543.162	.	
Fraudule...	Fraudule...	SBC	Schwarz'...	3170.616	.	
Fraudule...	Fraudule...	ASE	Average ...	0.049202	0.055845	
Fraudule...	Fraudule...	MAX	Maximu...	0.985593	0.999822	
Fraudule...	Fraudule...	DIV	Divisor fo...	5994	4002	
Fraudule...	Fraudule...	NOBS	Sum of F...	2997	2001	
Fraudule...	Fraudule...	RASE	Root Ave...	0.221815	0.236316	
Fraudule...	Fraudule...	SSE	Sum of S...	294.9164	223.4927	
Fraudule...	Fraudule...	SUMW	Sum of C...	5994	4002	
Fraudule...	Fraudule...	FPE	Final Pre...	0.058985	.	
Fraudule...	Fraudule...	MSE	Mean Sq...	0.054093	0.055845	
Fraudule...	Fraudule...	RFPE	Root Fin...	0.242867	.	
Fraudule...	Fraudule...	RMSE	Root Mea...	0.23258	0.236316	
Fraudule...	Fraudule...	AVERR	Average ...	0.167027	0.206924	
Fraudule...	Fraudule...	ERR	Error Fun...	1001.162	828.1108	
Fraudule...	Fraudule...	MISC	Misclassi...	0.060727	0.062969	
Fraudule...	Fraudule...	WRON...	Number ...	182	126	

2:26 PM
2/22/2023

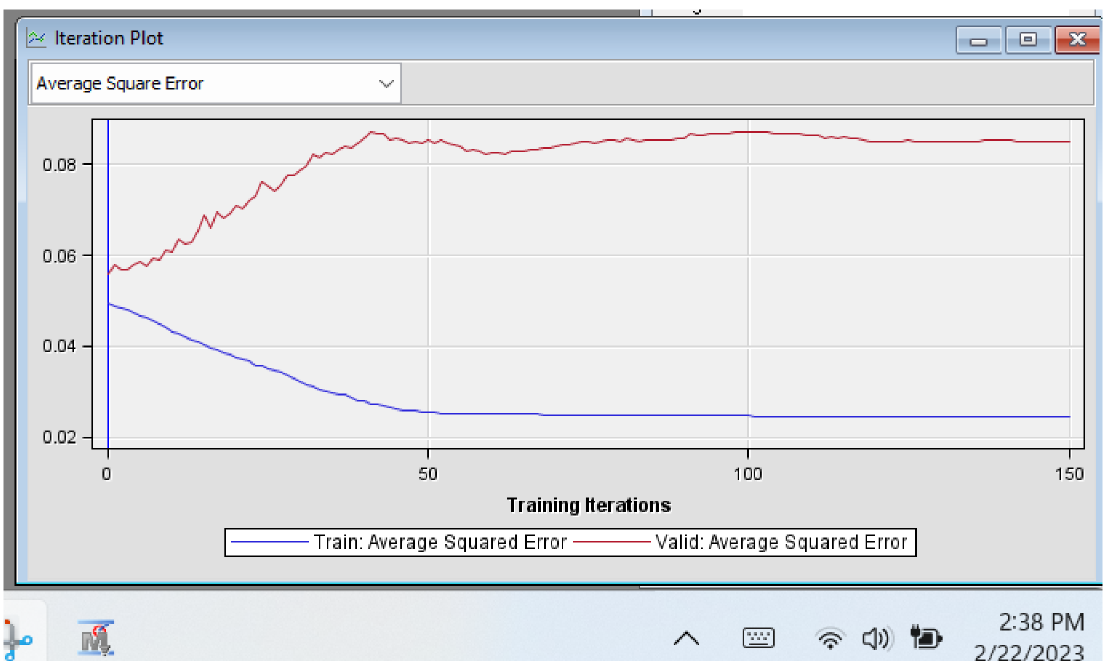
The first step is 15%, and the lift is 3.566. The result of the cumulation lift is higher than the cumulative lift of the model with three hidden models.

Figure 13: Cumulative Lift for the number of hidden units six and iteration 150 for the Multilayer Perception model neural network.



The training iteration plot results show that this model trains very quickly, and over the 100 iterations, the average square error goes straight to 150.

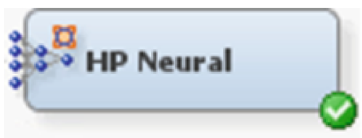
Figure 14: Iteration Plot for the number of hidden units six and iteration 150 for the Multilayer Perception model neural network.



Automatically Generate a Neural Network

The neural network has a property to create with few architectural considerations, including the AutoNeural and DMNeural nodes. The AutoNeural node has simple architectures with single, block, funnel, and cascade layers.

The DMNeural node mainly used for the target variable is binary or interval. It uses the nonlinear model to solve nonlinear estimation problems, reduce computing time, and find globally optimal solutions. The DMNeural node uses each of the eight action functions to choose the best. For example, the combination function will default to IDENT for a binary target and LOGIST for an interval target, and the node requires at least two input variables.



The HPNeural node has excellent performance for a large amount of data stores by minimizing the amount of data movement, and its unique properties are parallel processing and line memory. HPNeural property set automatically as input (s) and target(s) might be interval, binary, or nominal. In addition, the model handles missing values that the model may need to be addressed.

Explaining a Neural Network

A neural network is a complex and robust tool; however, most companies need help explaining how it works, which is challenging to understand. A decision tree may be described as a neural network—first action MLP neural network with six hidden units because it performed the result for the claim fraud dataset excellently. MLP neural network with Metadata node and decision tree. Drag and drop Metadata in the Utility tab to diagram the workplace, then connect the Neural network node to the Metadata node. Moreover, add a Decision tree node in the Model tab.

Figure 15: Neural network with decision tree node and Metadata node

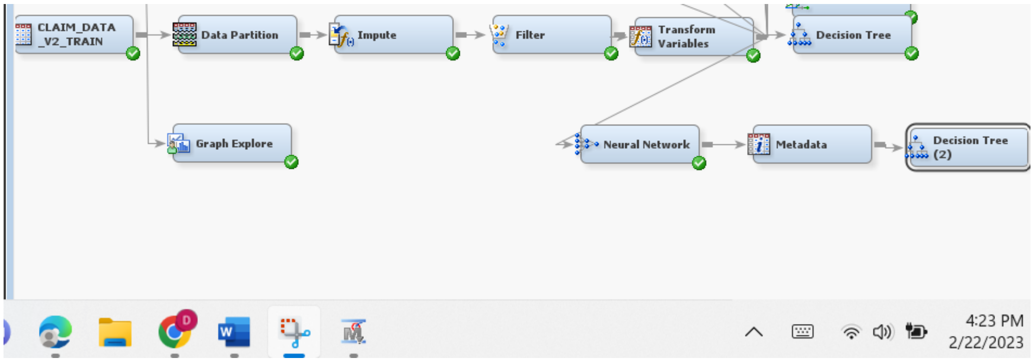
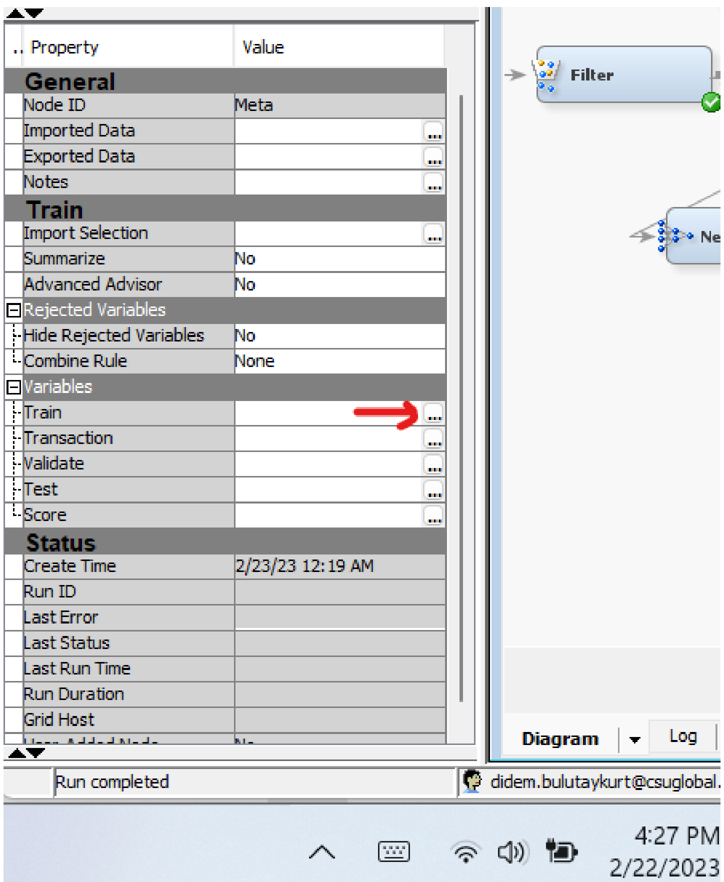
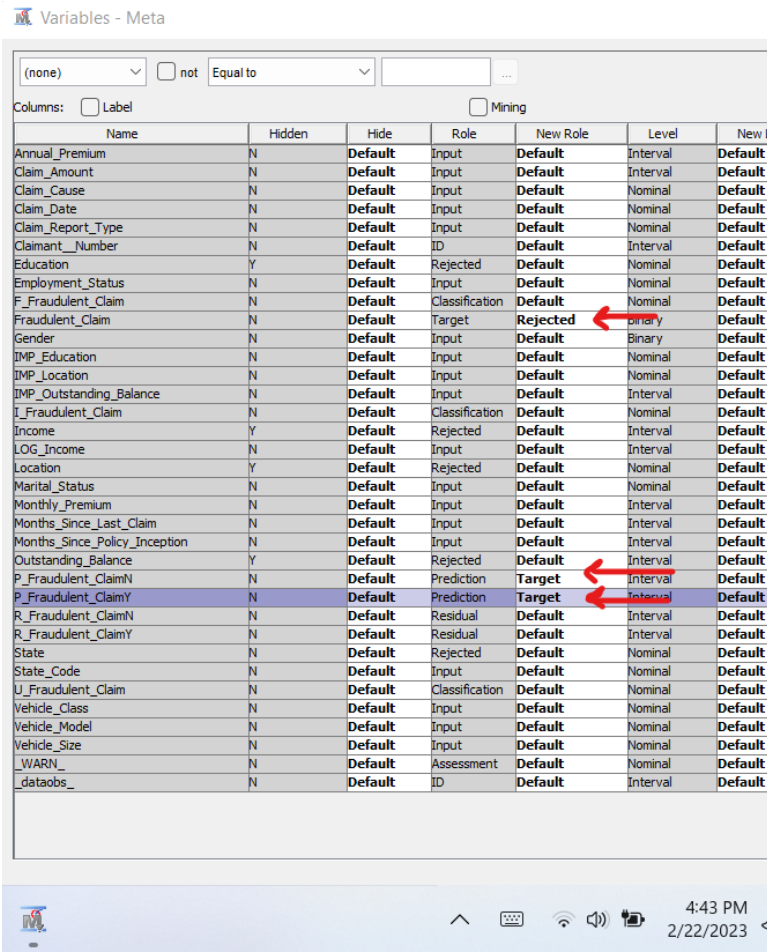


Figure 16: Update Train data set with Metadata node property



Click to Train ellipse to reject target variable fraud_calim, then use the variable generated by the neural network node. P_Fraudlent_ClaimN is the probability that the claim is not fraudulent, and P_Fraudulent_ClaimY is the probability of value that the claim is fraudulent. Set both target variables.

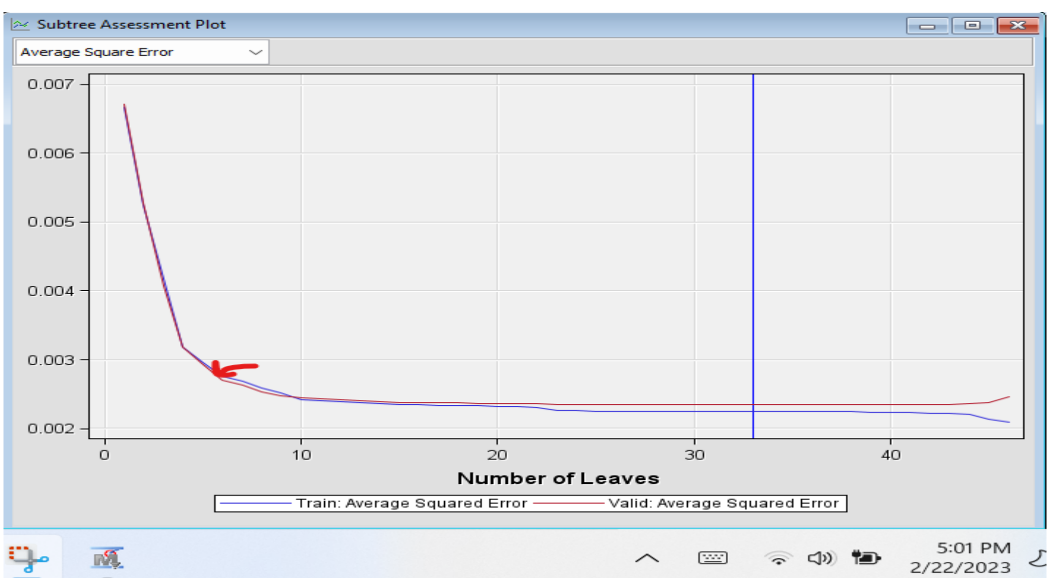
Figure 17: Metadata node Train property



Name	Hidden	Hide	Role	New Role	Level	New l
Annual_Premium	N	Default	Input	Default	Interval	Default
Claim_Amount	N	Default	Input	Default	Interval	Default
Claim_Cause	N	Default	Input	Default	Nominal	Default
Claim_Date	N	Default	Input	Default	Nominal	Default
Claim_Report_Type	N	Default	Input	Default	Nominal	Default
Claimant__Number	N	Default	ID	Default	Interval	Default
Education	Y	Default	Rejected	Default	Nominal	Default
Employment_Status	N	Default	Input	Default	Nominal	Default
F_Fraudulent_Claim	N	Default	Classification	Default	Nominal	Default
Fraudulent_Claim	N	Default	Target	Rejected	Binary	Default
Gender	N	Default	Input	Default	Binary	Default
IMP_Education	N	Default	Input	Default	Nominal	Default
IMP_Location	N	Default	Input	Default	Nominal	Default
IMP_Outstanding_Balance	N	Default	Input	Default	Interval	Default
I_Fraudulent_Claim	N	Default	Classification	Default	Nominal	Default
Income	Y	Default	Rejected	Default	Interval	Default
LOG_Income	N	Default	Input	Default	Interval	Default
Location	Y	Default	Rejected	Default	Nominal	Default
Marital_Status	N	Default	Input	Default	Nominal	Default
Monthly_Premium	N	Default	Input	Default	Interval	Default
Months_Since_Last_Claim	N	Default	Input	Default	Interval	Default
Months_Since_Policy_Inception	N	Default	Input	Default	Interval	Default
Outstanding_Balance	Y	Default	Rejected	Default	Interval	Default
P_Fraudulent_ClaimN	N	Default	Prediction	Target	Interval	Default
P_Fraudulent_ClaimY	N	Default	Prediction	Target	Interval	Default
R_Fraudulent_ClaimN	N	Default	Residual	Default	Interval	Default
R_Fraudulent_ClaimY	N	Default	Residual	Default	Interval	Default
State	N	Default	Rejected	Default	Nominal	Default
State_Code	N	Default	Input	Default	Nominal	Default
U_Fraudulent_Claim	N	Default	Classification	Default	Nominal	Default
Vehicle_Class	N	Default	Input	Default	Nominal	Default
Vehicle_Model	N	Default	Input	Default	Nominal	Default
Vehicle_Size	N	Default	Input	Default	Nominal	Default
WARN	N	Default	Assessment	Default	Nominal	Default
dataobs	N	Default	ID	Default	Interval	Default

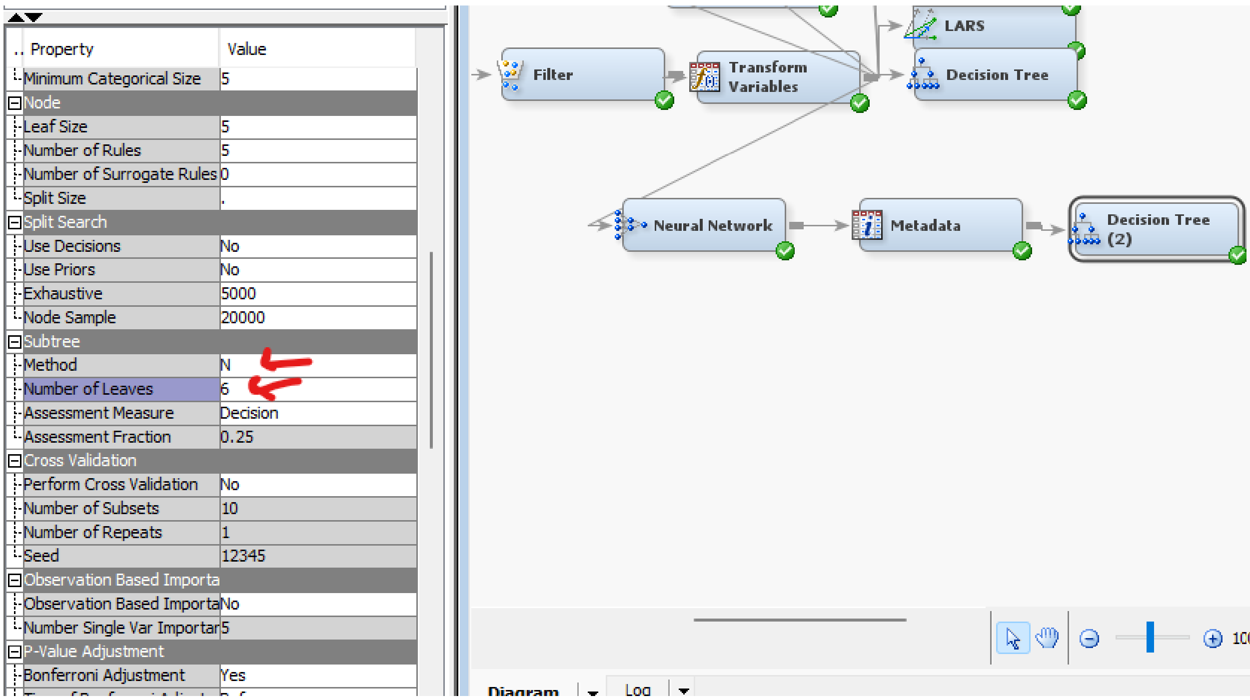
The result of the subtree assessment plot shows 33 leaves produced; after the sixth leaf, the remaining leaves do not have a significant impact.

Figure 18: Subtree assessment plot



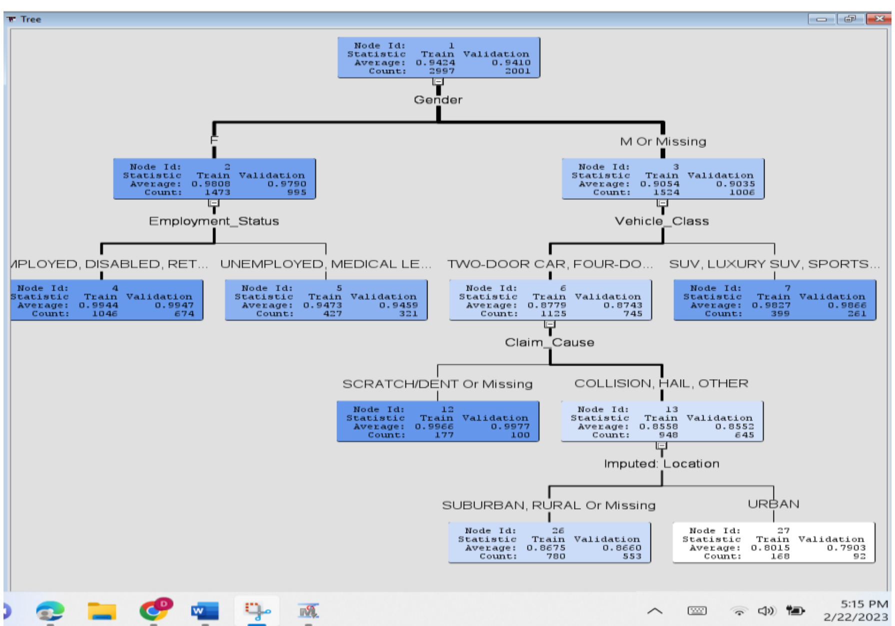
Let us look at how the Decision tree can be explained and closely approximates the significant factors for the neural network—Set Method property of the Decision tree node N and the number of leaves to 6.

Figure 19: Decision tree properties



The result of the decision tree output shows the most significant input variable on the neural network. The input variables are Gender, Employment_Status, Vehicle_Class, and Claim_Cause.

Figure 20: Decision tree output



Thus, darker color tells high Rcall, lighter color tells low Rcall, and to understand which variables had a significance as explain neural network model result. That way, it is possible to identify which variables should be removed from further analysis because they are not significant enough to support the cost of their inclusion.

<u>Model Comparison and Scoring</u>

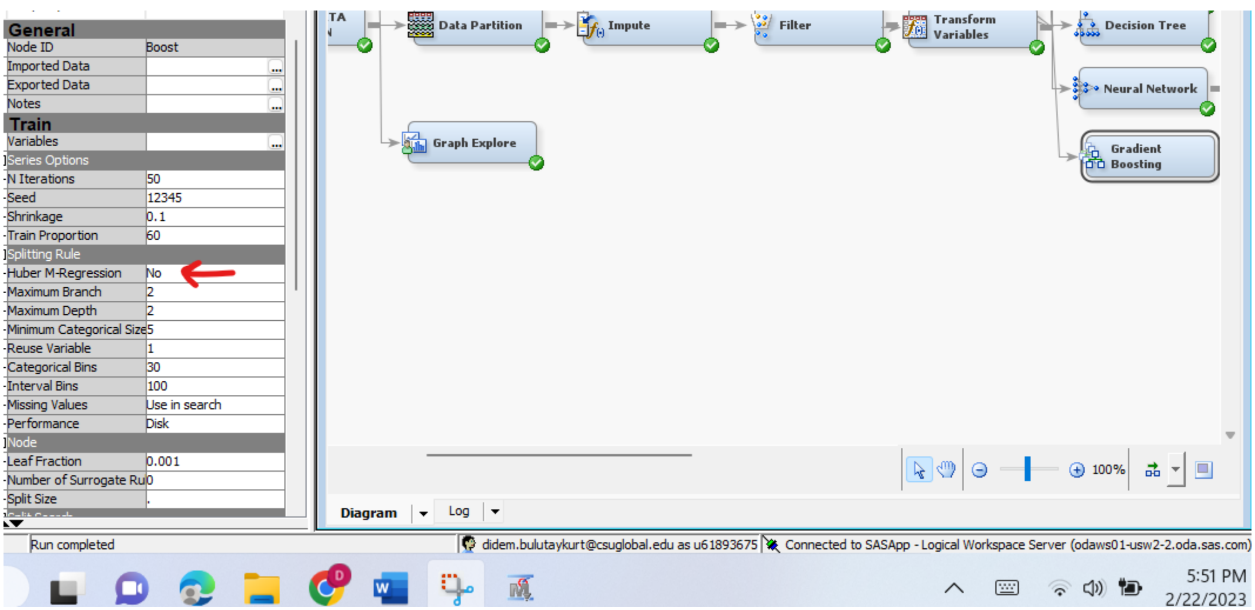
The Big 3 in the predictive analytics list are regression, decision tree, and neural network. In this chapter, I will work on a method to develop a model and compare results.

Gradient Boosting

The Gradient boosting node prepares the decision tree and regression algorithms for a large amount of data to produce a model as the combined technique to produce results for each technique. It may handle outliers and missing values better than decision trees or regression analysis. The model considers multiple algorithms; one of the best-known is XGBoost. That is designed to solve speed with parallel construction—Gradient boosting uses interval, nominal, and binary targets. If the target is an interval, the Huber M-Regression property should set the No; the square error function will be used because the Huber M-Regression loss function is less sensitive to outliers.

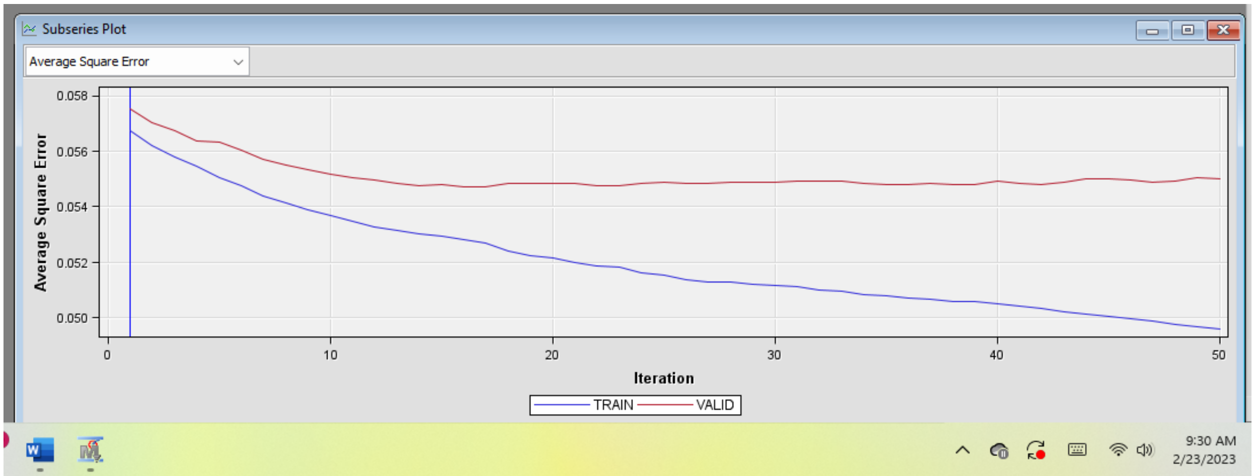
I will apply the claim fraud dataset to the Gradient boosting node in the Model tab.

Figure 21: Gradient boosting node and properties



The result average square error of a Gradient boosting model shows 50 leaves produced and that any iteration did not improve the average square error. However, the average square error increased during the 50 iterations.

Figure 22: Result of Subseries Plot Gradient boosting.



The variable Importance window shows the list of claim fraud and observation-based variable importance. The list of essential variables are Vehicle_Class, Gender, and Employment_Status; those are the most significant impacts on a gradient boosting model.

Figure 23: Variable Importance of Gradient Boosting Node

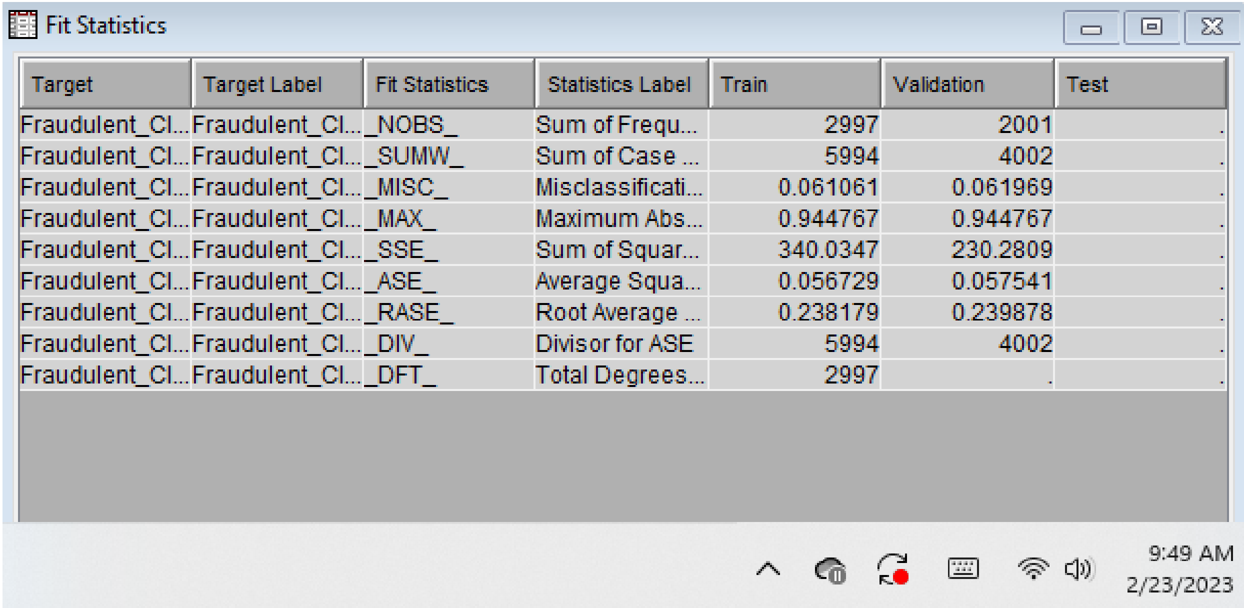
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Vehicle_Class	Vehicle_Class	1	1	0.923656	0.923656
Gender	Gender	1	0.975533	1	1.02508
Employment_St...	Employment_St...	1	0.54825	0.611162	1.114751
Claim_Report_T...	Claim_Report_T...	0	0	0	.
Claim_Date	Claim_Date	0	0	0	.
Claim_Amount	Claim_Amount	0	0	0	.
Monthly_Premium	Monthly_Premium	0	0	0	.

9:41 AM
2/23/2023

The result of the average square error claim insurance dataset with the Gradient boosting model is significantly higher than the result of the neural network model.

Figure 24: Fit Statistic of Gradient boosting node



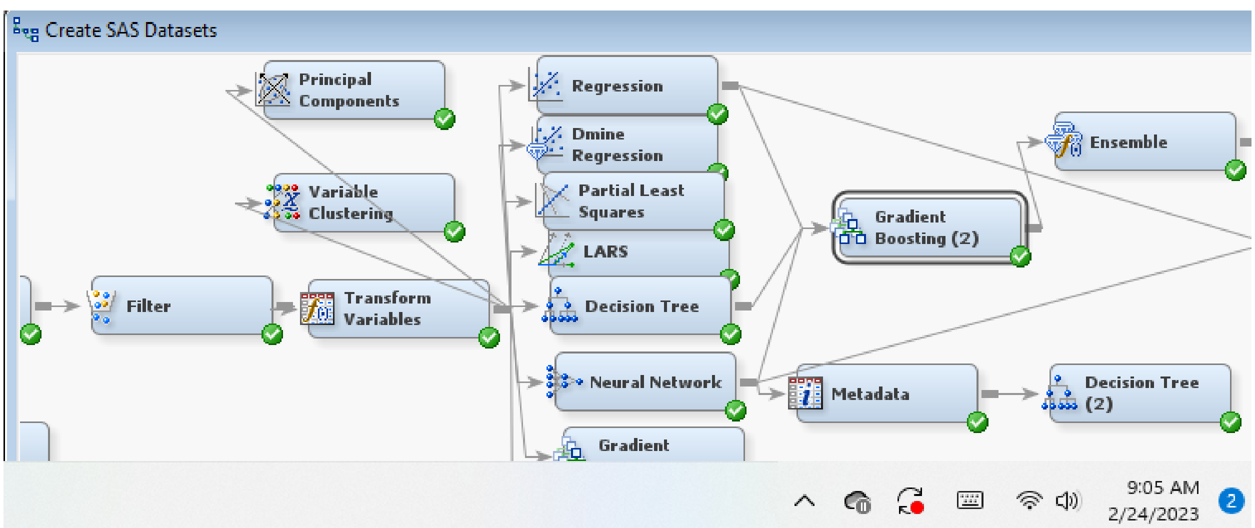
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Cl...	Fraudulent_Cl...	_NOBS_	Sum of Frequ...	2997	2001	.
Fraudulent_Cl...	Fraudulent_Cl...	_SUMW_	Sum of Case ...	5994	4002	.
Fraudulent_Cl...	Fraudulent_Cl...	_MISC_	Misclassificati...	0.061061	0.061969	.
Fraudulent_Cl...	Fraudulent_Cl...	_MAX_	Maximum Abs...	0.944767	0.944767	.
Fraudulent_Cl...	Fraudulent_Cl...	_SSE_	Sum of Squar...	340.0347	230.2809	.
Fraudulent_Cl...	Fraudulent_Cl...	_ASE_	Average Squa...	0.056729	0.057541	.
Fraudulent_Cl...	Fraudulent_Cl...	_RASE_	Root Average ...	0.238179	0.239878	.
Fraudulent_Cl...	Fraudulent_Cl...	_DIV_	Divisor for ASE	5994	4002	.
Fraudulent_Cl...	Fraudulent_Cl...	_DFT_	Total Degrees...	2997	.	.

Ensemble Models

The ensemble node is used independently of each other with the same target variable and predicts the interval targets or probability of nominal or binary targets. An ensemble model best fits the decision tree, neural network, and regression model. Gradient boosting may also work. Additionally, random forests cannot serve as input to a unit. So, the best way to compare the individual model's results to each other and the Ensemble should perform to the respective models. Finally, I will apply the claim insurance dataset to the ensemble model.

Figure 25: Ensemble node and properties



I used a regression model, decision tree, and neural network in combination with the Gradient boosting, so the results improved the average square error; however, the models combined were not identical as each represented the best of their group. Thus, the ensemble model is thought to be an improvement.

Figure 26: Fit Statistics of Ensemble node

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Cl...	Fraudulent_Cl...	_ASE_	Average Squa...	0.056729	0.057541	.
Fraudulent_Cl...	Fraudulent_Cl...	_DIV_	Divisor for ASE	5994	4002	.
Fraudulent_Cl...	Fraudulent_Cl...	_MAX_	Maximum Abs...	0.944767	0.944767	.
Fraudulent_Cl...	Fraudulent_Cl...	_NOBS_	Sum of Frequ...	2997	2001	.
Fraudulent_Cl...	Fraudulent_Cl...	_RASE_	Root Average ...	0.238179	0.239878	.
Fraudulent_Cl...	Fraudulent_Cl...	_SSE_	Sum of Squar...	340.0347	230.2809	.
Fraudulent_Cl...	Fraudulent_Cl...	_DISF_	Frequency of ...	2997	2001	.
Fraudulent_Cl...	Fraudulent_Cl...	_MISC_	Misclassificati...	0.061061	0.061969	.
Fraudulent_Cl...	Fraudulent_Cl...	_WRONG_	Number of Wr...	183	124	.

9:03 AM
2/24/2023

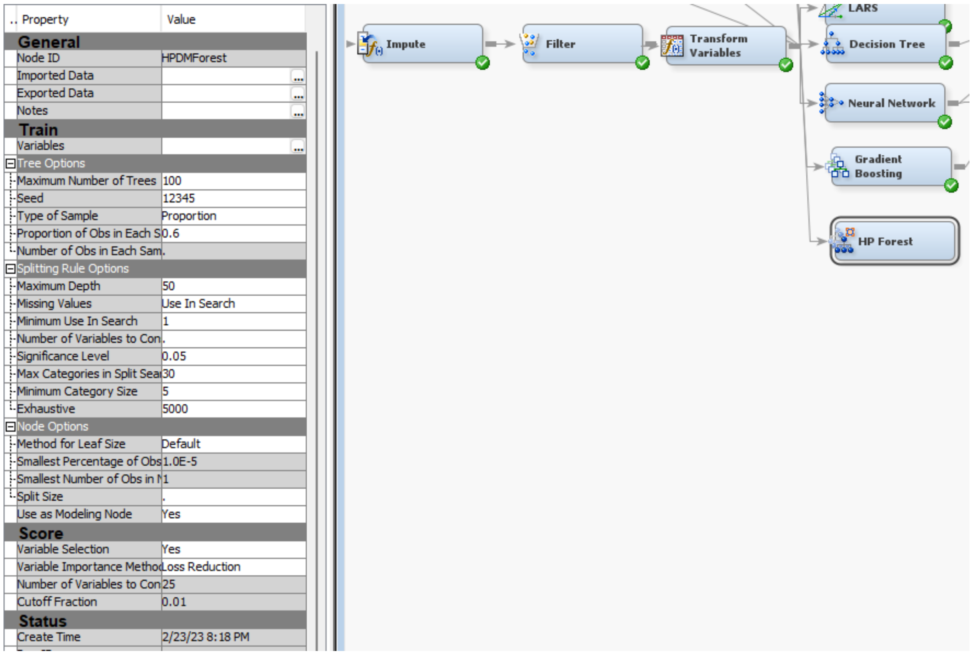
Random Forest

The Random forest performs multiple decision trees to support regression and classification trees. Combining numerous trees into the forest aims for a more accurate prediction than a single decision tree.

The HP Forest node works with big data sets that use the average of many trees to create a single tree model. The best property of random forest works with regression and classification trees, meaning the target can be binary, nominal, or interval. The worst thing about the random forest model is that it requires more trees to improve accuracy as it increases run times, especially when applying large datasets.

Drag and drop HP Forest in the HPDM tab to connect Transform Variables of the claim fraud dataset.

Figure 27: HP Forest node and properties



Property	Value
General	
Node ID	HPDMForest
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Tree Options	
Maximum Number of Trees	100
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each S	0.6
Number of Obs in Each Sam	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Con	.
Significance Level	0.05
Max Categories in Split Sea	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest Percentage of Obs	1.0E-5
Smallest Number of Obs in N	1
Split Size	.
Use as Modeling Node	Yes
Score	
Variable Selection	Yes
Variable Importance Metho	Loss Reduction
Number of Variables to Con	25
Cutoff Fraction	0.01
Status	
Create Time	2/23/23 8:18 PM

The result of the average square error of 0.055155 is lower than the MLP neural network result of 0.055845.

Figure 28: Fit Statistics result of random forest node

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent...	Fraudulent...	_ASE_	Average Sq...	0.054133	0.055155	.
Fraudulent...	Fraudulent...	_DIV_	Divisor for A...	5994	4002	.
Fraudulent...	Fraudulent...	_MAX_	Maximum A...	0.964003	0.950304	.
Fraudulent...	Fraudulent...	_NOBS_	Sum of Fre...	2997	2001	.
Fraudulent...	Fraudulent...	_RASE_	Root Avera...	0.232664	0.234851	.
Fraudulent...	Fraudulent...	_SSE_	Sum of Squ...	324.4704	220.7307	.
Fraudulent...	Fraudulent...	_DISF_	Frequency ...	2997	2001	.
Fraudulent...	Fraudulent...	_MISC_	Misclassific...	0.061061	0.061969	.
Fraudulent...	Fraudulent...	_WRONG_	Number of ...	183	124	.

12:29 PM
2/23/2023

The Iteration History window shows the iterations of the tree structure that generated the result and how quickly the model builds to the final result.

Figure 29: Iteration History on Random Forest node

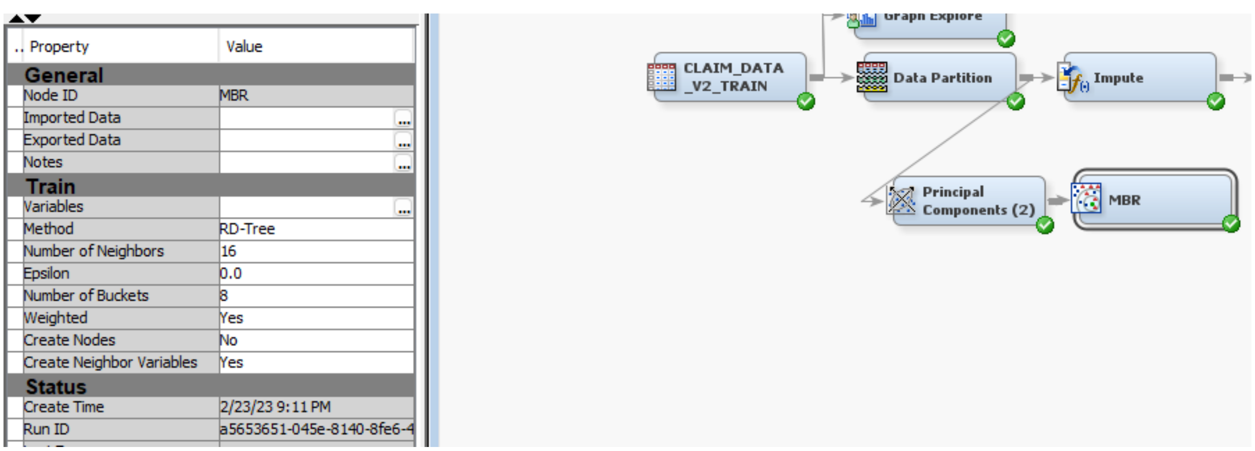
Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (Out of Bag)	Average Square Error (Validate)	Misclassification Rate (Train)	Misclassification Rate (Out of Bag)	Misclassification Rate (Validate)	Log Loss (Train)	Log Loss (Out of Bag)	Log Loss (Validate)
1	5	0.0542	0.0538	0.0550	0.0611	0.0601	0.062	0.200	0.211	0.199
2	8	0.0545	0.0542	0.0554	0.0611	0.0594	0.062	0.203	0.215	0.206
3	11	0.0548	0.0555	0.0556	0.0611	0.0609	0.062	0.205	0.212	0.208
4	13	0.0551	0.0545	0.0559	0.0611	0.0594	0.062	0.210	0.210	0.212
5	16	0.0549	0.0533	0.0557	0.0611	0.0584	0.062	0.207	0.205	0.210
6	18	0.0550	0.0544	0.0558	0.0611	0.0595	0.062	0.209	0.209	0.212
7	21	0.0549	0.0547	0.0558	0.0611	0.0601	0.062	0.207	0.208	0.211
8	23	0.0550	0.0548	0.0559	0.0611	0.0602	0.062	0.209	0.209	0.212
9	26	0.0549	0.0552	0.0557	0.0611	0.0608	0.062	0.208	0.210	0.211
10	29	0.0550	0.0550	0.0559	0.0611	0.0605	0.062	0.209	0.211	0.213
11	31	0.0550	0.0552	0.0560	0.0611	0.0607	0.062	0.209	0.211	0.213
12	33	0.0550	0.0556	0.0559	0.0611	0.0612	0.062	0.209	0.213	0.213
13	37	0.0548	0.0554	0.0558	0.0611	0.0612	0.062	0.208	0.212	0.212
14	46	0.0546	0.0553	0.0557	0.0611	0.0611	0.062	0.206	0.211	0.211
15	54	0.0544	0.0551	0.0555	0.0611	0.0611	0.062	0.205	0.209	0.209
16	64	0.0540	0.0548	0.0552	0.0611	0.0611	0.062	0.202	0.207	0.207
17	69	0.0539	0.0546	0.0551	0.0611	0.0611	0.062	0.201	0.205	0.206
18	76	0.0538	0.0545	0.0550	0.0611	0.0611	0.062	0.200	0.204	0.205
19	79	0.0538	0.0545	0.0550	0.0611	0.0611	0.062	0.200	0.204	0.205
20	85	0.0537	0.0545	0.0550	0.0611	0.0611	0.062	0.199	0.204	0.205

Memory-Based Reasoning

The model uses the k-nearest neighbor algorithm to produce an observed classification method to compare cases to previous cases and apply historical data to build records like current cases. The k-nearest neighbor algorithm calculates the distance that Euclidean distance. The input variables must be numeric, so the categorical variable must be transformed into numeric values. It might be necessary to reduce the number of categorical variables. A memory-based reasoning node contains only one target variable. The target variable might be nominal, binary, or interval.

I will apply a memory-based reasoning node to the claim fraud dataset, and first, I will use the Principal Components node for utilization and then connect.

Figure 30: Principal Components node and MBR node



To set the default method RD-Tree, as a result, shows an average square error of 0.056764 and a high number of average square errors between the train and validation dataset that could be better as the ensemble model results.

Figure 31: Fit Statistics on MBR node

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Cl...	Fraudulent_Cl...	_NW_	Number of Est...	19	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_NOBS_	Sum of Frequ...	2998	2003	.
Fraudulent_Cl...	Fraudulent_Cl...	_SUMW_	Sum of Case ...	5996	4006	.
Fraudulent_Cl...	Fraudulent_Cl...	_DFT_	Total Degrees...	2998	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_DFM_	Model Degree...	19	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_DFE_	Degrees of Fr...	2979	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_ASE_	Average Squa...	0.051709	0.056764	.
Fraudulent_Cl...	Fraudulent_Cl...	_RASE_	Root Average ...	0.227396	0.238253	.
Fraudulent_Cl...	Fraudulent_Cl...	_DIV_	Divisor for ASE	5996	4006	.
Fraudulent_Cl...	Fraudulent_Cl...	_SSE_	Sum of Squar...	310.0469	227.3984	.
Fraudulent_Cl...	Fraudulent_Cl...	_MSE_	Mean Square...	0.052039	0.056764	.
Fraudulent_Cl...	Fraudulent_Cl...	_RMSE_	Root Mean Sq...	0.22812	0.238253	.
Fraudulent_Cl...	Fraudulent_Cl...	_AVERR_	Average Error ...	0.179289	0.250345	.
Fraudulent_Cl...	Fraudulent_Cl...	_ERR_	Error Function	1075.019	1002.881	.
Fraudulent_Cl...	Fraudulent_Cl...	_MAX_	Maximum Abs...	0.9375	1	.
Fraudulent_Cl...	Fraudulent_Cl...	_FPE_	Final Predictio...	0.052369	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_RFPE_	Root Final Pre...	0.228842	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_AIC_	Akaike's Infor...	1113.019	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_SBC_	Schwarz's Bay...	1227.128	.	.
Fraudulent_Cl...	Fraudulent_Cl...	_MISC_	Misclassificati...	0.061041	0.061907	.
Fraudulent_Cl...	Fraudulent_Cl...	_WRONG_	Number of Wr...	183	124	.

1:23 PM
2/23/2023

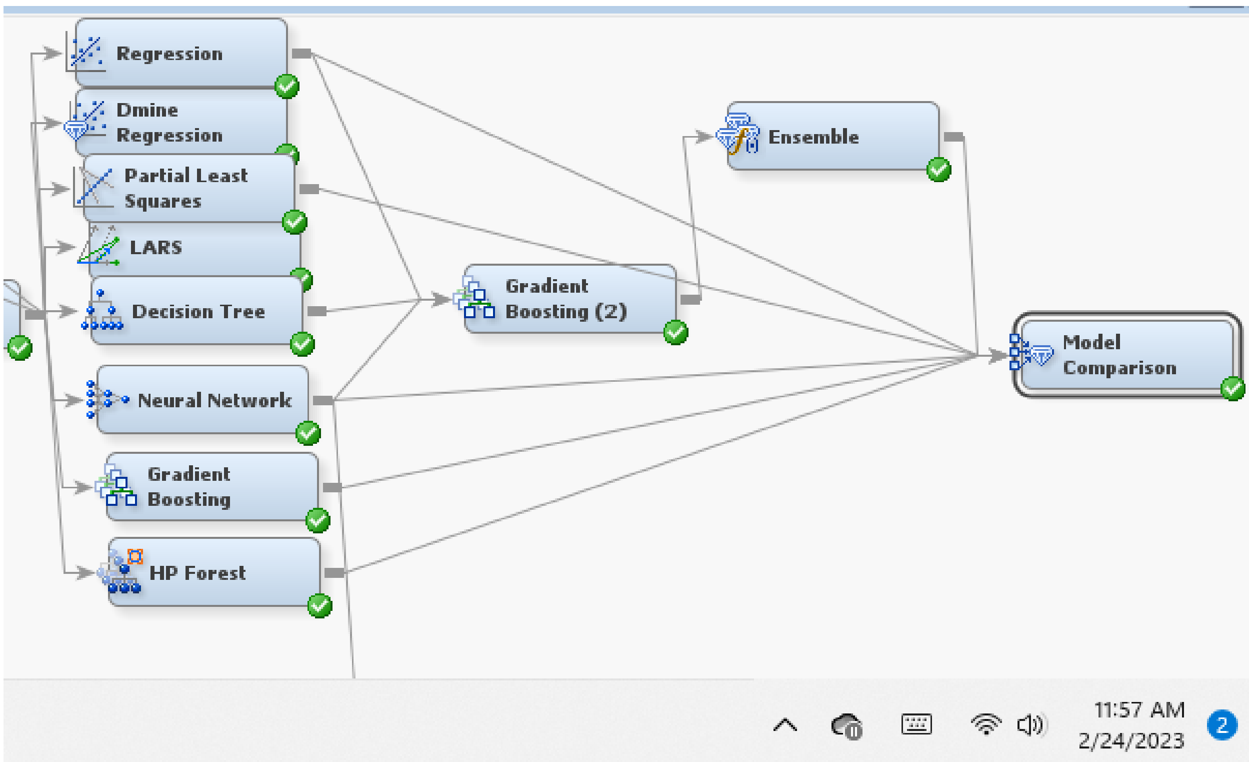
Two-Stage Model

The two-stage node processes two target variables at the same time. One of the target variables is a class variable; the other is an interval variable that is generally accurate, the value related to the level of the class variable. The default function builds a categorical prediction variable from the class target and then uses it to model the interval target.

Comparing Predictive Models

If there are two or more predictive models, they should be able to compare them to find which model best fits. The Model Comparison node compares models and predictions from other models like regression, decision trees, or neural networks. For example, I applied a claim fraud dataset to a few models. Drag and drop the Model Comparison node in the Assess tab on the diagram workplace.

Figure 32: Model Comparison node



Evaluating Fit Statistic

SAS Enterprise Miner has 14 different statistical results to compare model performance. I will explain a few of them with the claim fraud dataset result I applied.

The misclassification rate is among the most valuable statistics results, especially when the target value is binary. When comparing models, the best result is the lowest misclassification rate. For example, as a result of the misclassification rate for the claim fraud dataset, the decision tree, PLS, HP Forest, Gradient Boosting, Ensemble, and regression have the same misclassification rate, except the Neural network has a higher value than others, which does not help decide best- fit model however we can eliminate Neural network.

Figure 33: The result of the misclassification rate on the Model Comparison node

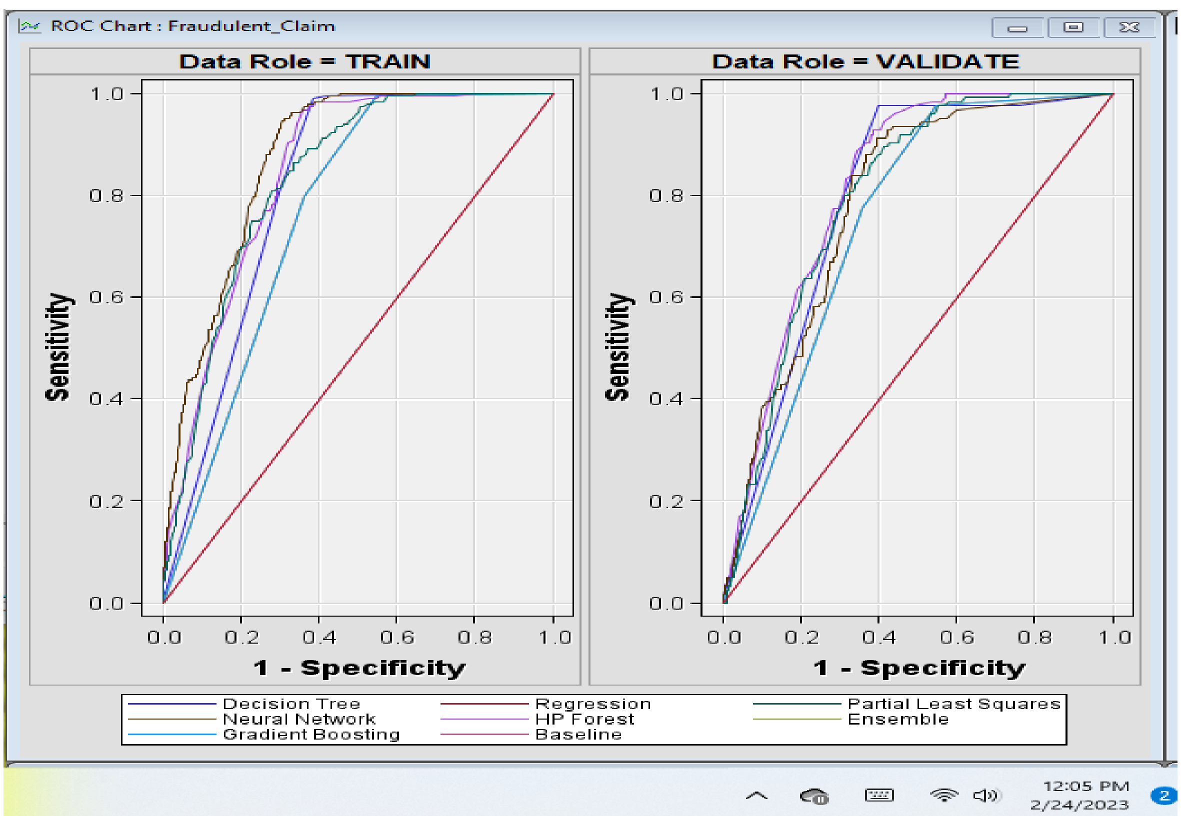
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree	Decision Tree	0.061969	0.052381	0.061061	0.053507
	PLS	Partial Least Squares	0.061969	0.052995	0.061061	0.054664
	HPDMForest	HP Forest	0.061969	0.054133	0.061061	0.055155
	Boost	Gradient Boosting	0.061969	0.056729	0.061061	0.057541
	Ensmbl	Ensemble	0.061969	0.056729	0.061061	0.057541
	Reg	Regression	0.061969	0.057333	0.061061	0.058130
	Neural	Neural Network	0.062969	0.049202	0.060727	0.055845

12:05 PM
2/24/2023

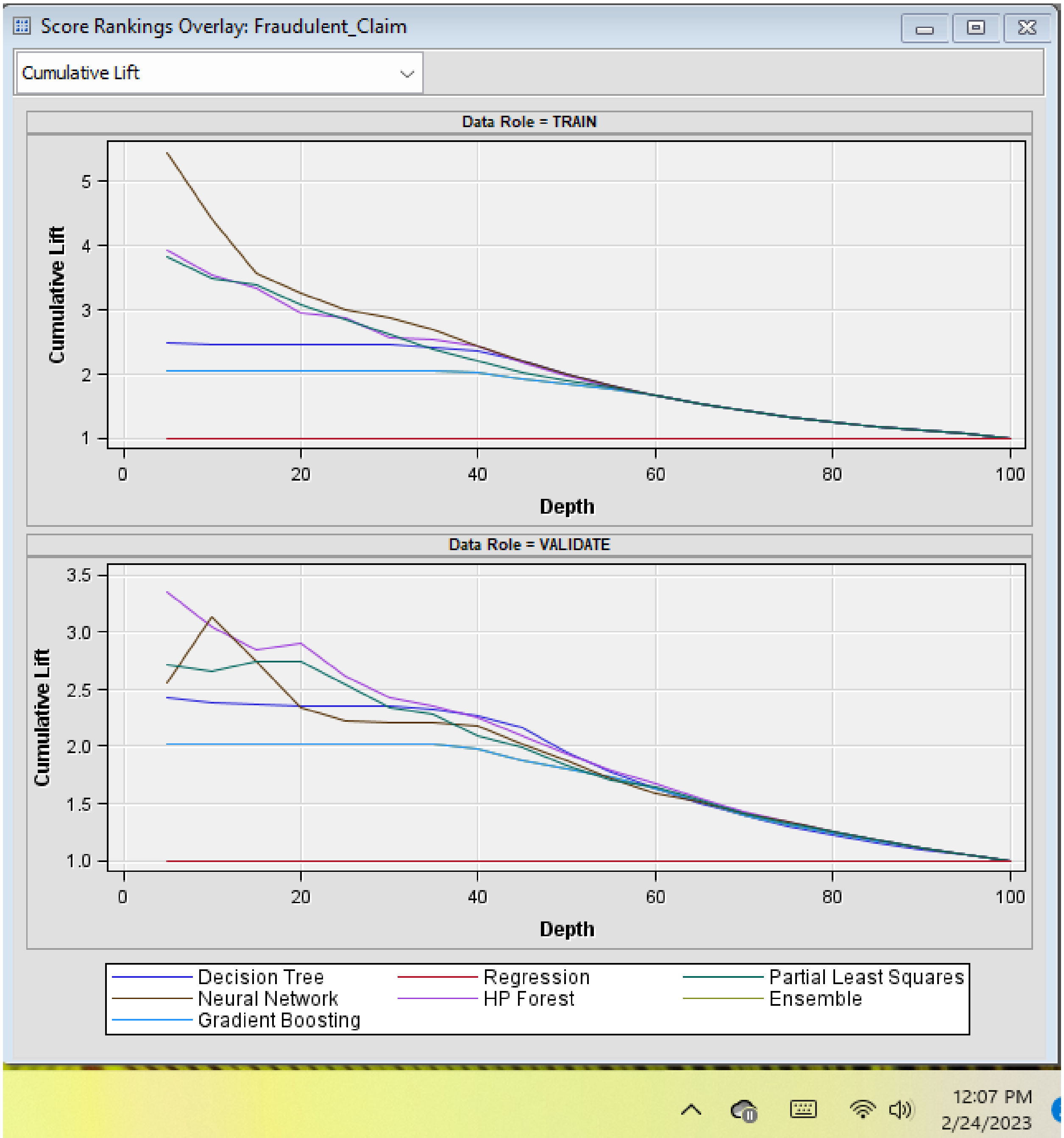
The receiver operating characteristic curve displays sensitivity as the y-axis and specificity as the x-axis of the ROC curve. Under the curve is C-statistics (concordance), which shows the goodness of fit for the binary outcomes. The model's large area under the curve best fits when comparing the models. If the ROC index is smaller than six, it is weak. If the ROC index is higher than seven, the index is considered to be strong. The result of the ROC graph for the claim fraud dataset best predicts models is a decision tree.

Figure 34: ROC curve into Model Comparison node



The cumulative lift measure is used to estimate the performance of random model guessing. The x-axis shows the result of the percentage of the overall data. Comparing the models showed that the highest number of lifts was more robust than the model. As a result of claim fraud, cumulative charge shows the highest number of lifts at 2.86 and 20% depth from the HP Forest model on the validation dataset.

Figure 35: Cumulative Lift window model comparison



Conclusion

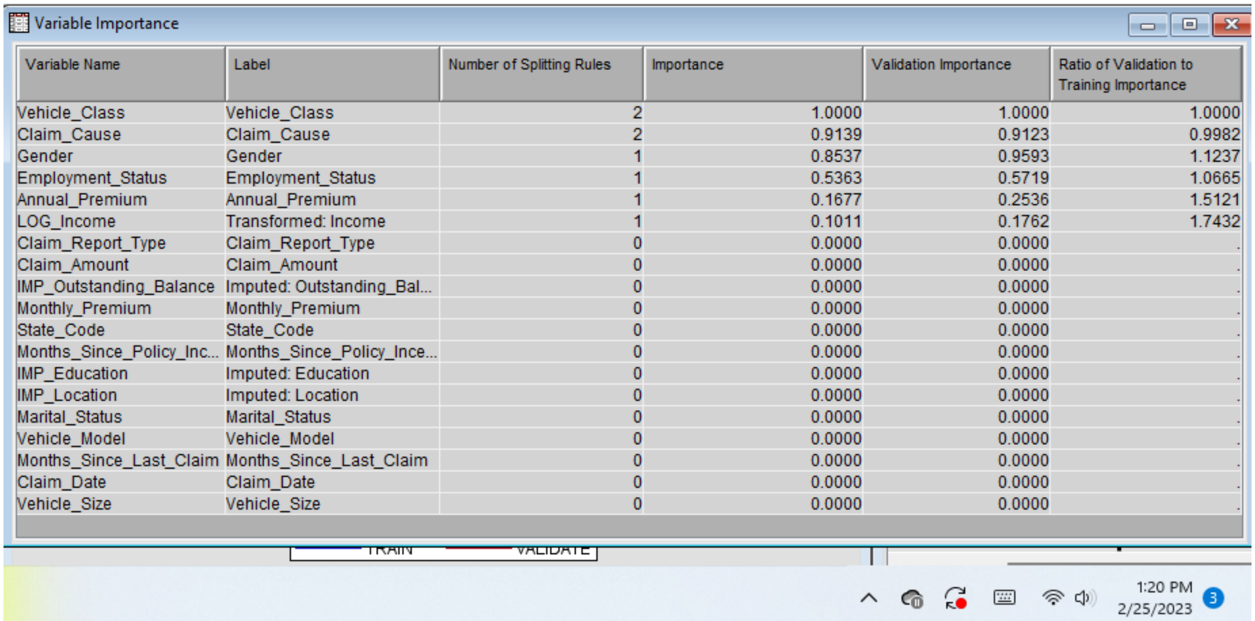
"After all, it is the best-fit model that should be used to analyze current business activity. The most common statistics for evaluating predictive models include the misclassification rate, average squared error, ROC index, and cumulative lift. " (McCarthy,2022)

The result of the average square error is more able to trust measure in these cases. The great for model comparison with the lowest errors is to appraise the best-fit model. Figure 33 shows the average square error in the validation dataset; the lowest error is the decision tree of 0.053507.

Model Name	Average Square error	Depth	Cumulative lift
Generalized Linear Model	0.0555546	15%	3.38
Generalized Linear Model(target layer combination and activation set Linear)	0.059704	-	-
Multilayer Perception model (tree hidden layer)	0.056747	15%	3.12
Multilayer Perception model (three hidden units and 150 iterations)	0.055845	15%	3.566
Gradient Boosting	0.057541	15%	2.04
Ensemble	0.057541	15%	2.04
HP Forest	0.05155	15%	2.83
MBR	0.056764	15%	3.39
Decision Tree	0.053507	15%	2.45

The result of the decision tree's important variable is that the auto insurance claim fraud case's most predictive variable for the target variable is Fraudulent_Claim. The most significant variables for the future Vehicle_Class, Claim_Cause, Gender, Employment_Status, Annual_Premium, and transformed income.

Figure 36: Variable Importance window by Decision Tree Model.


Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Vehicle_Class	Vehicle_Class	2	1.0000	1.0000	1.0000
Claim_Cause	Claim_Cause	2	0.9139	0.9123	0.9982
Gender	Gender	1	0.8537	0.9593	1.1237
Employment_Status	Employment_Status	1	0.5363	0.5719	1.0665
Annual_Premium	Annual_Premium	1	0.1677	0.2536	1.5121
LOG_Income	Transformed: Income	1	0.1011	0.1762	1.7432
Claim_Report_Type	Claim_Report_Type	0	0.0000	0.0000	.
Claim_Amount	Claim_Amount	0	0.0000	0.0000	.
IMP_Outstanding_Balance	Imputed: Outstanding_Bal...	0	0.0000	0.0000	.
Monthly_Premium	Monthly_Premium	0	0.0000	0.0000	.
State_Code	State_Code	0	0.0000	0.0000	.
Months_Since_Policy_Inc...	Months_Since_Policy_Ince...	0	0.0000	0.0000	.
IMP_Education	Imputed: Education	0	0.0000	0.0000	.
IMP_Location	Imputed: Location	0	0.0000	0.0000	.
Marital_Status	Marital_Status	0	0.0000	0.0000	.
Vehicle_Model	Vehicle_Model	0	0.0000	0.0000	.
Months_Since_Last_Claim	Months_Since_Last_Claim	0	0.0000	0.0000	.
Claim_Date	Claim_Date	0	0.0000	0.0000	.
Vehicle_Size	Vehicle_Size	0	0.0000	0.0000	.

Using Historical Data to predict the future with Score node

The main aim is to find the best fit for the historical data analysis. The SAS Enterprise Mine has a Score node applying an existing predictive model to new transaction data to measure probability or anticipate value for a target variable outcome. The probability result will explain the prediction if the target variable is binary or nominal. The anticipated variable will be calculated if the target is an interval. In this process, two inputs are essential to the Score node. Process one is the scored dataset; the other predictive model connects the Score node.

Figure 37: Creating a score dataset.

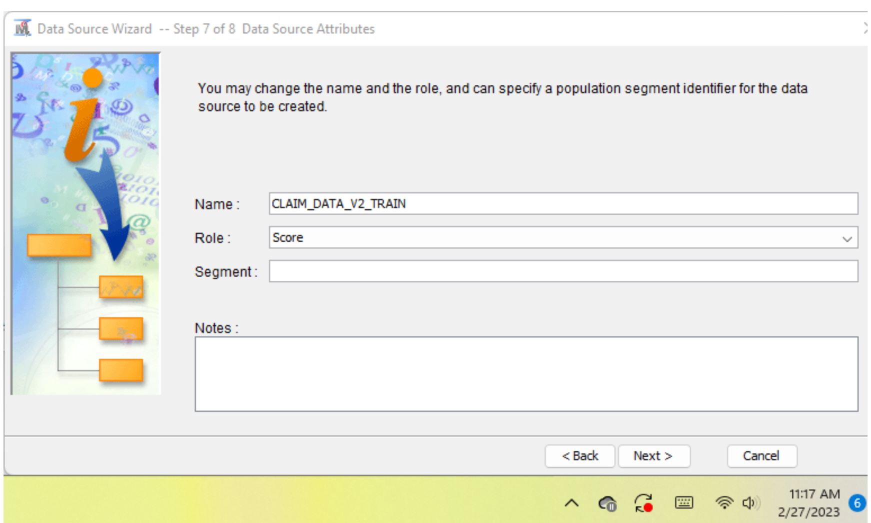
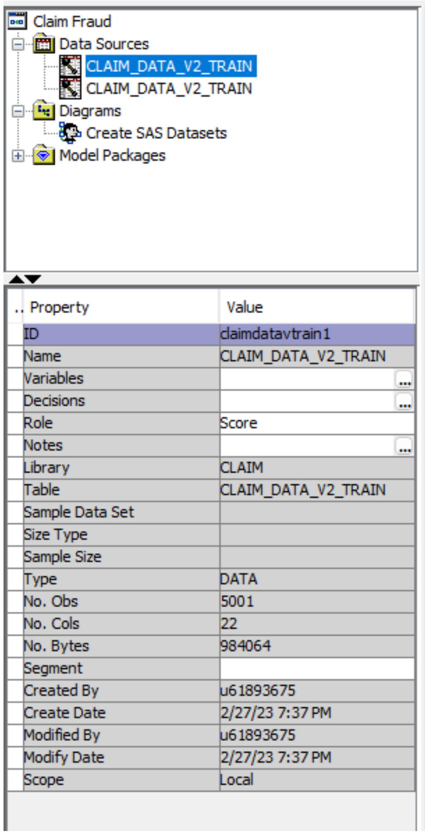
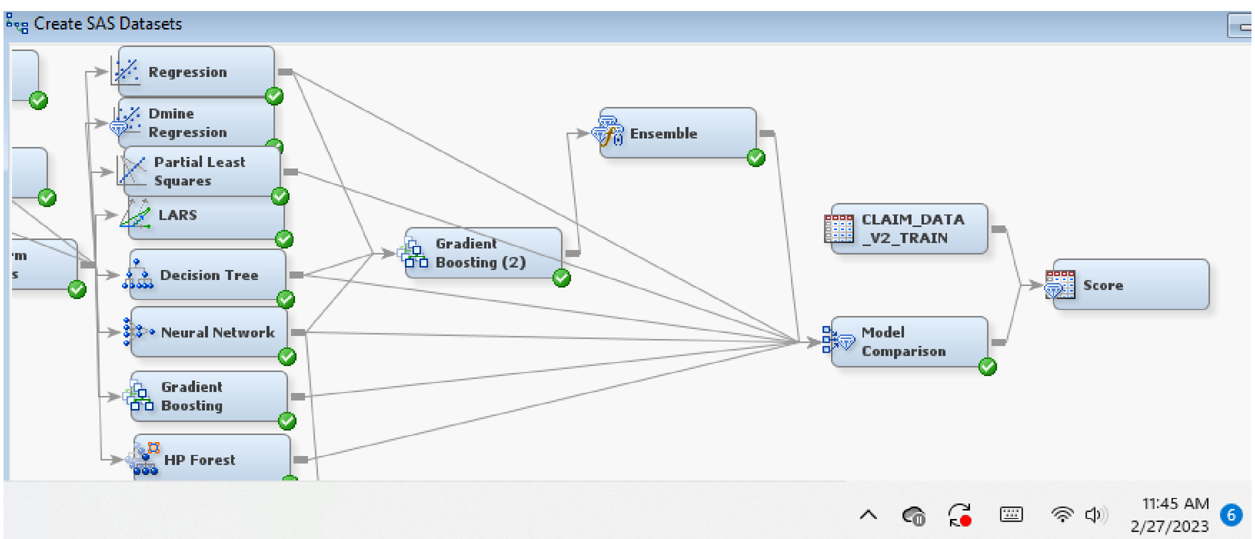


Figure 38: CLAIM_DATA_V2_TRAIN Score Role data property



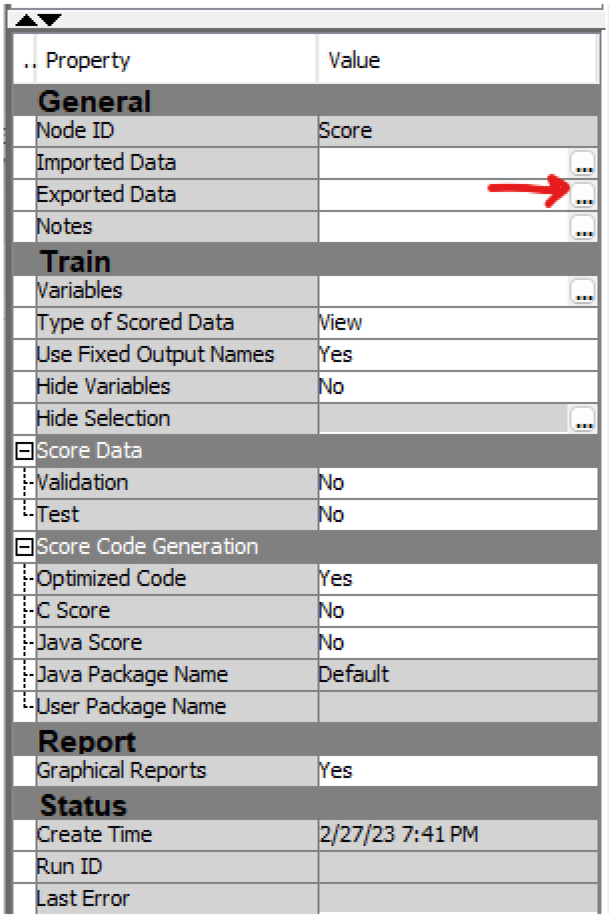
After the scored dataset's role, drag and drop the Score node in the Assess tab to diagram the workplace. Then, connect a role of the score dataset and model comparison to the Score node.

Figure 39: Scoring node



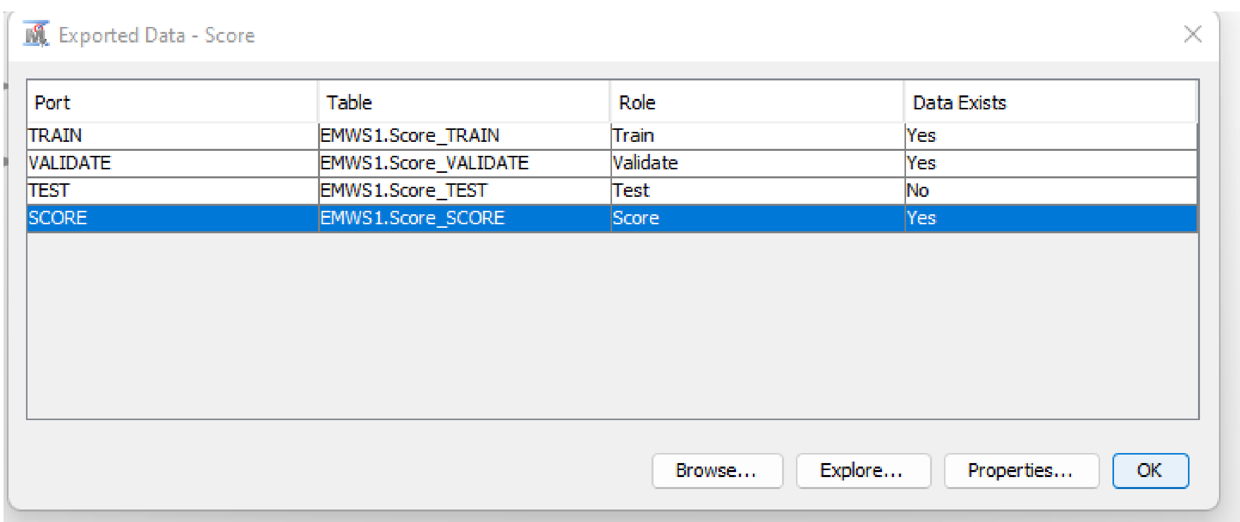
The next step is to click the Exported Data ellipse from the Score node properties.

Figure 40: Score node properties



A new window will show all the available datasets; choose the SCORE dataset and select Explore.

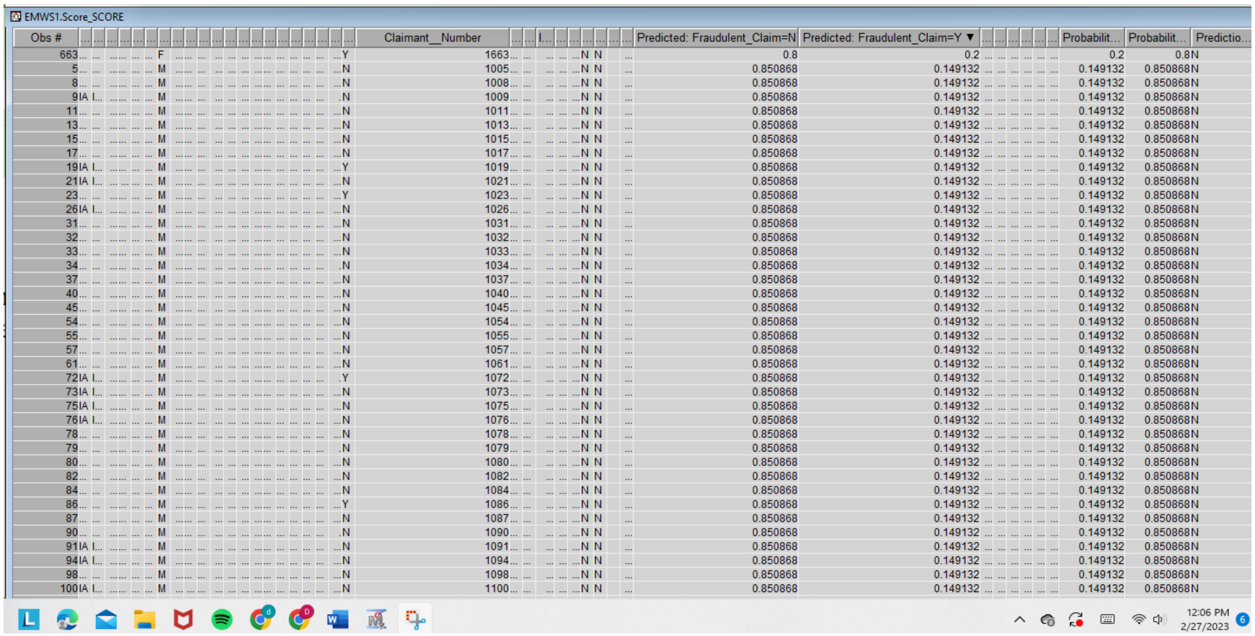
Figure 41: Exported data.



Analyzing and Reporting Results

The result of exploring the score dataset output of the auto insurance claim dataset's binary target variable has two predictions: Predicted Fraud_Claim=N and Predicted Fraud_Claim=Y.

Figure 42: Explore the score dataset. The output shows the probabilities by Claimant_Number.



EMWS1.Score_SCORE

Obs #	...	...	...	Claimant__Number	...	...	Predicted: Fraudulent_Claim=N	Predicted: Fraudulent_Claim=Y ▼	Probabilit...	Probabilit...	Predictio...
663	...	F	...Y	1663	...N	N	0.8	0.2	0.2	0.8	N
5	...	M	...N	1005	...N	N	0.850868	0.149132	0.149132	0.850868	N
8	...	M	...N	1008	...N	N	0.850868	0.149132	0.149132	0.850868	N
9	IA I...	M	.N	1009	...N	N	0.850868	0.149132	0.149132	0.850868	N
11	...	M	...N	1011	...N	N	0.850868	0.149132	0.149132	0.850868	N
13	...	M	...N	1013	...N	N	0.850868	0.149132	0.149132	0.850868	N
15	...	M	...N	1015	...N	N	0.850868	0.149132	0.149132	0.850868	N
17	...	M	...N	1017	...N	N	0.850868	0.149132	0.149132	0.850868	N
19	IA I...	M	...Y	1019	...N	N	0.850868	0.149132	0.149132	0.850868	N
21	IA I...	M	...N	1021	...N	N	0.850868	0.149132	0.149132	0.850868	N
23	...	M	...Y	1023	...N	N	0.850868	0.149132	0.149132	0.850868	N
26	IA I...	M	...N	1026	...N	N	0.850868	0.149132	0.149132	0.850868	N
31	...	M	...N	1031	...N	N	0.850868	0.149132	0.149132	0.850868	N
32	...	M	...N	1032	...N	N	0.850868	0.149132	0.149132	0.850868	N
33	...	M	...N	1033	...N	N	0.850868	0.149132	0.149132	0.850868	N
34	...	M	.N	1034	...N	N	0.850868	0.149132	0.149132	0.850868	N
37	...	M	...N	1037	...N	N	0.850868	0.149132	0.149132	0.850868	N
40	...	M	...N	1040	...N	N	0.850868	0.149132	0.149132	0.850868	N
45	...	M	...N	1045	...N	N	0.850868	0.149132	0.149132	0.850868	N
54	...	M	...N	1054	...N	N	0.850868	0.149132	0.149132	0.850868	N
55	...	M	...N	1055	...N	N	0.850868	0.149132	0.149132	0.850868	N
57	...	M	...N	1057	...N	N	0.850868	0.149132	0.149132	0.850868	N
61	...	M	...N	1061	...N	N	0.850868	0.149132	0.149132	0.850868	N
72	IA I...	M	.Y	1072	...N	N	0.850868	0.149132	0.149132	0.850868	N
73	IA I...	M	...N	1073	...N	N	0.850868	0.149132	0.149132	0.850868	N
75	IA I...	M	...N	1075	...N	N	0.850868	0.149132	0.149132	0.850868	N
76	IA I...	M	...N	1076	...N	N	0.850868	0.149132	0.149132	0.850868	N
78	...	M	...N	1078	...N	N	0.850868	0.149132	0.149132	0.850868	N
79	...	M	.N	1079	...N	N	0.850868	0.149132	0.149132	0.850868	N
80	...	M	...N	1080	...N	N	0.850868	0.149132	0.149132	0.850868	N
82	...	M	...N	1082	...N	N	0.850868	0.149132	0.149132	0.850868	N
84	...	M	...N	1084	...N	N	0.850868	0.149132	0.149132	0.850868	N
86	...	M	...Y	1086	...N	N	0.850868	0.149132	0.149132	0.850868	N
87	...	M	...N	1087	...N	N	0.850868	0.149132	0.149132	0.850868	N
90	...	M	.N	1090	...N	N	0.850868	0.149132	0.149132	0.850868	N
91	IA I...	M	...N	1091	...N	N	0.850868	0.149132	0.149132	0.850868	N
94	IA I...	M	...N	1094	...N	N	0.850868	0.149132	0.149132	0.850868	N
98	...	M	...N	1098	...N	N	0.850868	0.149132	0.149132	0.850868	N
100	IA I...	M	...N	1100	...N	N	0.850868	0.149132	0.149132	0.850868	N

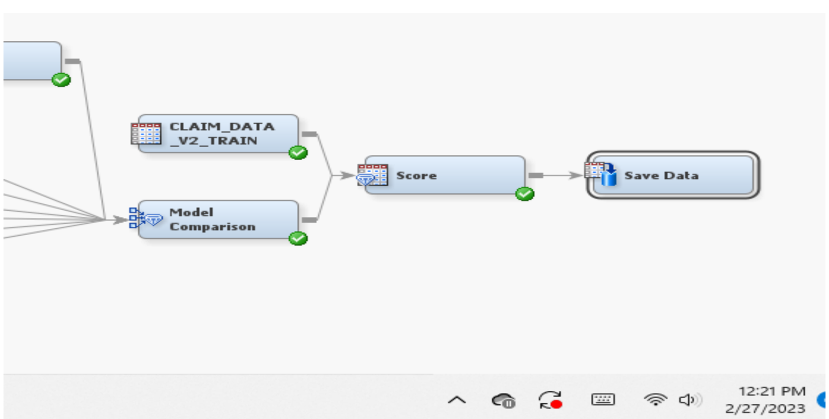
12:06 PM
2/27/2023

Column Predicted Fraudulent_Claim=Y sorts it shows the highest probability of fraud claim records. This type of claim is most likely fraudulent for further investigation. The dataset has a lower probability that the claim is fraudulent. Thus, the organization can utilize those results to make decisions.

Save Data Node

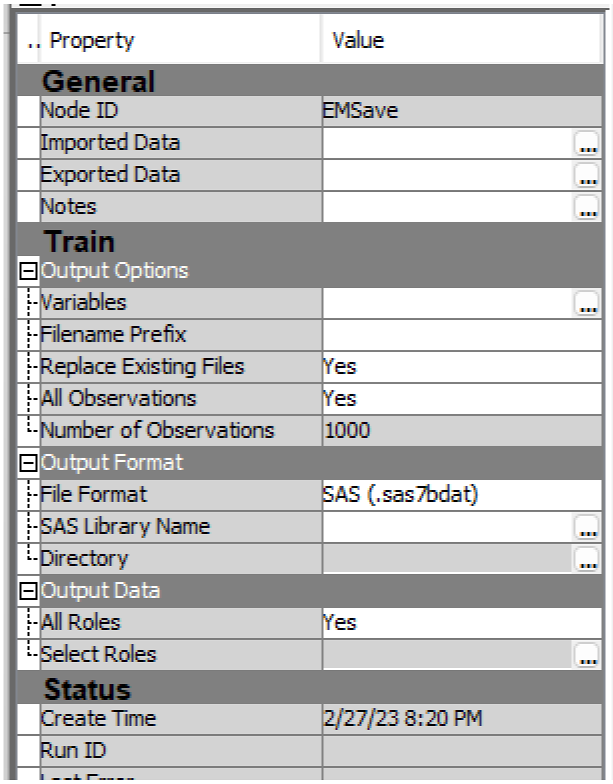
At the end of the predictive analysis, the SAS Enterprise Miner has a Save Data node to keep the dataset for future use. Drag and drop the Save data node in the Utility tab on the diagram workplace and connect the Score node.

Figure 43: Save Data node.



Two essential properties should be set: File Format and SAS Library Name.

Figure 44: Save Data properties.



Property	Value
General	
Node ID	EMSave
Imported Data	...
Exported Data	...
Notes	...
Train	
Output Options	
Variables	...
Filename Prefix	
Replace Existing Files	Yes
All Observations	Yes
Number of Observations	1000
Output Format	
File Format	SAS (.sas7bdat)
SAS Library Name	...
Directory	...
Output Data	
All Roles	Yes
Select Roles	...
Status	
Create Time	2/27/23 8:20 PM
Run ID	

Reporter Node

SAS Enterprise Miner has a Reporter node report of the entire model from the beginning of each subsequent node to the final node, as property default is reported in a PDF format. Drag and drop the Reporter node in the Utility tab on the diagram workplace.

Figure 45: Reporter node and properties.

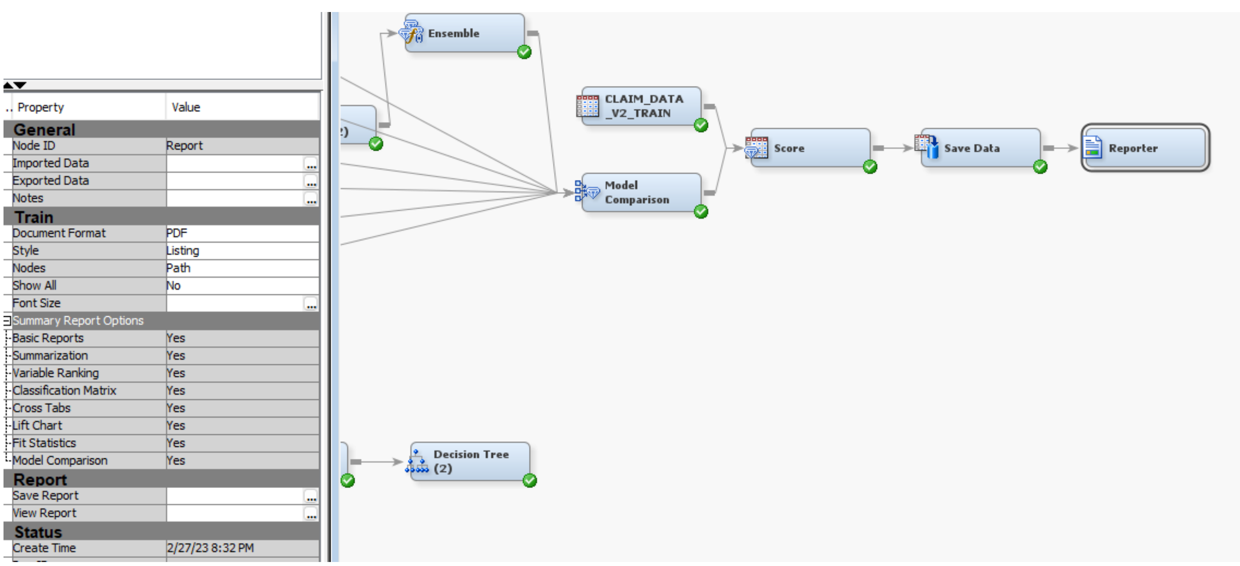


Figure 46: The Output of the Report node.

Output

```

                   Measurement    Frequency
Role                  Level         Count

ASSESS               BINARY            1
ASSESS               INTERVAL          2
ASSESS               NOMINAL           1
CLASSIFICATION       NOMINAL           3
ID                   INTERVAL          2
INPUT                BINARY            1
INPUT                INTERVAL          2
INPUT                NOMINAL           3
PREDICT              INTERVAL          4
REJECTED             INTERVAL          7
REJECTED             NOMINAL          10
RESIDUAL             INTERVAL          2
SEGMENT              NOMINAL           3
TARGET               BINARY            1




User       = u61893675
Date       = 20:34:47  27 February 2023
Project    = Claim Fraud
Diagram    = Create SAS Datasets

Start Node = Report
Node label = Save Data
Nodes      = PATH
Showall    = N

Format     = PDF
Graphics   = GIF
Style      = LISTING
```

Reference

Richard V. McCarthy, Mary M. McCarthy; Wendy Ceccucci, 2022. *Applying Predictive Analytics Finding Value in Data*. Second edition.

Donald Hebb, (1949). The Organization of Behavior.