ABSTRACT

Cardiovascular diseases are the number one cause of death globally, taking an average of 17.9 million lives each year, accounting for 31% of all deaths worldwide The purpose of our analysis was to determine key factors that are most closely indicative of a heart disease diagnosis using 11 variables provided by the dataset. Various decision trees, regression models, and neural networks were run for this analysis and a model comparison was conducted to determine which model was most predictive of the target. Using ASE scores and Roc Index values, we determined that the best model was the Backward Exclusion Regression Model. The output of this model revealed that Asymptomatic Chest Pain, raised Oldpeak levels, and age are factors that are most indicative of heart disease. From this analysis, we recommend that the best course of action to avoid heart disease is maintaining a healthy diet, regular exercise, and routine assessment if you have pre-existing risk factors.

Introduction:

Cardiovascular disease (CVD) is a term used to refer to the range of diseases affecting the heart and blood vessels. These include hypertension (high blood pressure), coronary heart disease (heart attack), cerebrovascular disease (stroke), and heart failure. CVDs are the number one cause of death globally, taking an average of 17.9 million lives each year, accounting for 31% of all deaths worldwide. While these conditions are expected in older populations, one-third of CVD deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

Workspace:

SAS Enterprise Miner is an advanced analytical data mining tool designed to quickly develop descriptive and predictive models by streamlining the data mining process. It assists analysts in identifying key relationships, recognizing patterns, and comparing data models.

SAS Models:

1. **Decision Trees:**
   - Maximal Tree
   - Decision Tree using Misclassification Rate Assessment
   - Decision Tree using Average Square Error Assessment
2. **Regression Models:**
   - Full Regression
   - Forward Regression
   - Backward Regression
   - Stepwise Regression
3. **Neural Networks:**
   - Neural Network with 3 hidden units and 100 iterations

- Neural Network with 3 hidden units and 50 iterations
- Neural Network with set upper and lower limits
- Neural Network with 4 hidden units
- Neural Network using variables selected by backward regression

Objective:

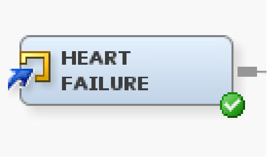
People with cardiovascular disease or who are at high cardiovascular risk need early detection and treatment to increase their chances of survival. The purpose of this analysis is to identify the key features of cardiovascular disease so that patients will receive immediate diagnosis and assistance.

Data Source:

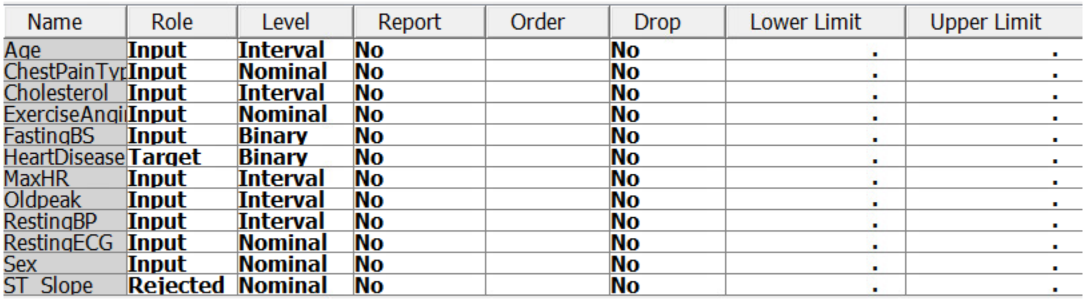
The data we will be using for this analysis is an excel file. Therefore, we must first import the file into our SAS Miner workspace.

Procedure:

1. Create an empty diagram called **Heart Failure.**
2. Select and drag the **File Import** node into the diagram from the sample tab.
3. Rename the node **Heart Failure.**
4. Select the **Heart Failure** node and click on Imported Data in the properties panel.
5. Locate and upload the data file.



Exploring the Data:



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
ChestPainTy	Input	Nominal	No		No	.	.
Cholesterol	Input	Interval	No		No	.	.
ExerciseAngi	Input	Nominal	No		No	.	.
FastingBS	Input	Binary	No		No	.	.
HeartDisease	Target	Binary	No		No	.	.
MaxHR	Input	Interval	No		No	.	.
Oldpeak	Input	Interval	No		No	.	.
RestingBP	Input	Interval	No		No	.	.
RestingECG	Input	Nominal	No		No	.	.
Sex	Input	Nominal	No		No	.	.
ST_Slope	Rejected	Nominal	No		No	.	.

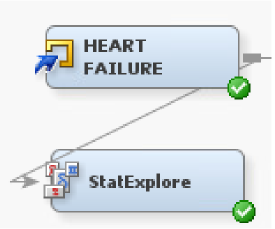
Target Variable: Heart Disease. This is a binary variable, with a positive diagnosis represented by 1 and a negative diagnosis represented by 0.

Rejected Variables: ST Slope. This measures the ST segment shift relative to increments in heart rate due to exercise. It is considered a more accurate ECG criterion for diagnosing heart disease. However, this variable is redundant, as ST Slope tests are conducted after CVDs are already detected. Therefore, it has been rejected from the dataset.

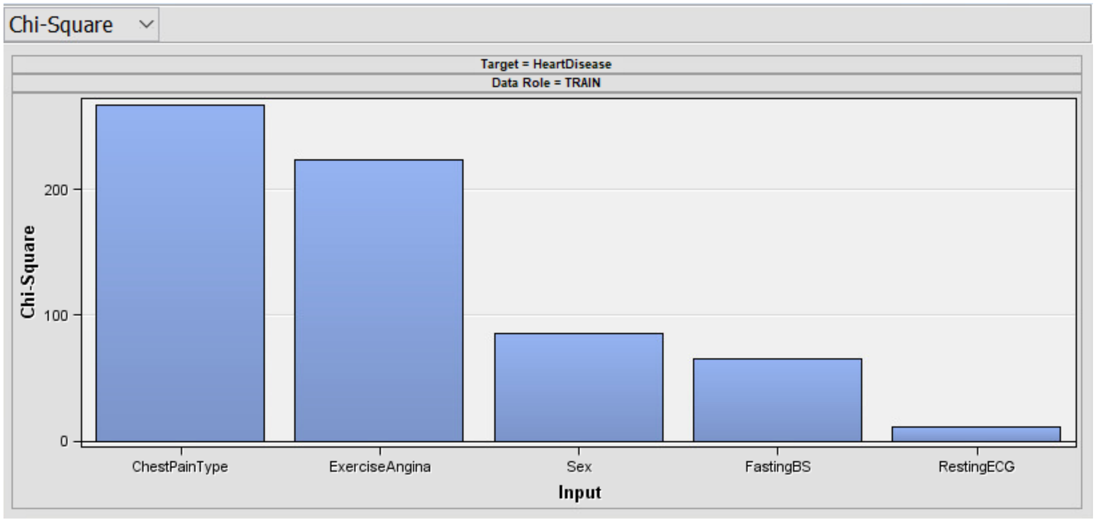
Accepted Variables:

1. **Age:** The age of studied patients in years, ranging from 28-77.
2. **Chest Pain Type:** Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NAP), Asymptomatic (ASY).
3. **Cholesterol:** Measures of serum cholesterol.
4. **Exercise Angina:** Chest pain induced by exercise.
5. **Fasting BS:** Fasting blood sugar levels. 0 indicates low levels and 1 indicates high levels.
6. **Max HR:** Maximum heart rate achieved.
7. **Oldpeak:** flat sections of an ECG. Elevation indicates a severe heart attack.
8. **Resting BP:** Resting blood pressure.
9. **Resting ECG:** Resting electrocardiogram results.
10. **Sex:** Male (M) or Female (F).

To explore the data in greater detail we selected the **StatExplore** node from the explore tab and connected it to the **Heart Failure** node.



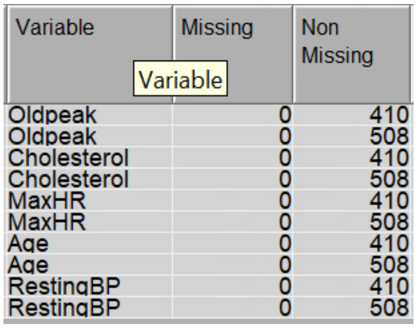
Select **StatExplore** and run the node. Open the results.



Chi-square testing is a statistical hypothesis testing method to observe the quality of fit between observed values and theoretically expected values. According to this Chi-Square Plot, the variables of ChestPainType, ExerciseAngina, Sex, FastingBS, and RestingEGC will be most significant in our analysis.

Missing Values:

Select View > Summary Statistics > Interval Variables from the results tab.



Variable	Missing	Non Missing
Oldpeak	0	410
Oldpeak	0	508
Cholesterol	0	410
Cholesterol	0	508
MaxHR	0	410
MaxHR	0	508
Age	0	410
Age	0	508
RestingBP	0	410
RestingBP	0	508

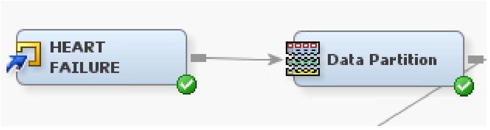
As indicated by this chart, there are no missing variables in our data that require replacement or imputation. Note that later in our analysis we will be replacing extreme values as Missing and giving them a new value using imputation.

Data Partitioning:

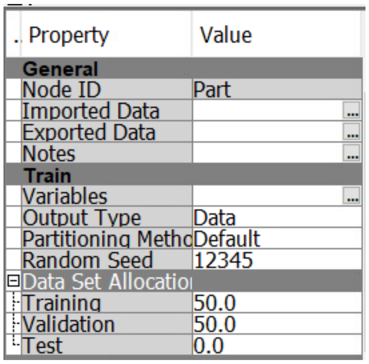
Splitting data, or data partitioning, is a standard procedure for honest model performance assessment when running predictive models. We will split our data into two parts: Training data (50%) which is used for fitting the data and Validation data (50%) which is used for monitoring and modifying the data to create better generalizations. Overfitting would be decreased by a larger dataset. If we are limited to the data in our existing dataset and unable to collect any more, we can use data augmentation to fictitiously enhance the size of our dataset. ***

Procedure:

1. Drag the **Data Partition** node from the sample tab into the diagram and connect it to your dataset.

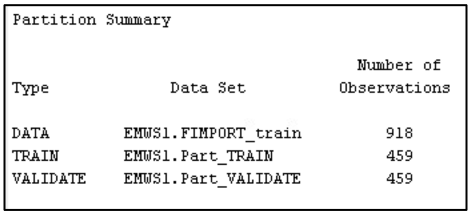


2. Allocate 50% to Training and 50% to Validation in the properties tab. Set Test to 0%.



Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Metho	Default
Random Seed	12345
⊟ Data Set Allocatio	
Training	50.0
Validation	50.0
Test	0.0

3. Run the node.



```
Partition Summary


                                        Number of
Type              Data Set            Observations

DATA        EMWS1.FIMPORT_train           918
TRAIN       EMWS1.Part_TRAIN              459
VALIDATE    EMWS1.Part_VALIDATE           459
```

Decision Trees

Decision trees are one of the best predictive modeling tools used. Input selection is conducted by a split search algorithm that rejects any variables with a log worth below 0.7. The complexity of decision trees is reduced by pruning so that the resulting tree only includes variables above the p-value threshold. The initial split is the Root Node, and the final splits are the Leaf Nodes. For all of our decision trees we will be using a two-branch mode as the majority of are variables are either binary or their values are split above and below a threshold.

Under this project, we have created three types of decision trees:

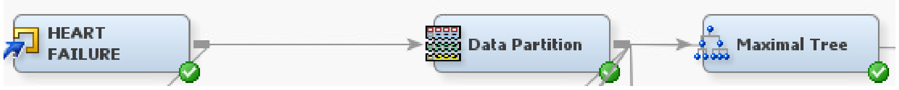
- Maximal Tree
- Misclassification Tree
- ASE Tree

Maximal Tree:

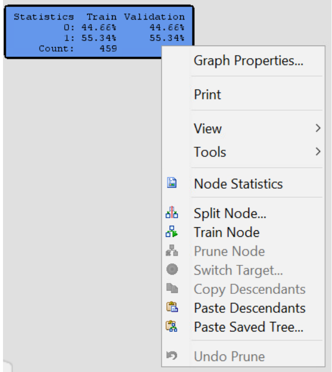
A maximal decision tree is one that has the maximal number of splits.

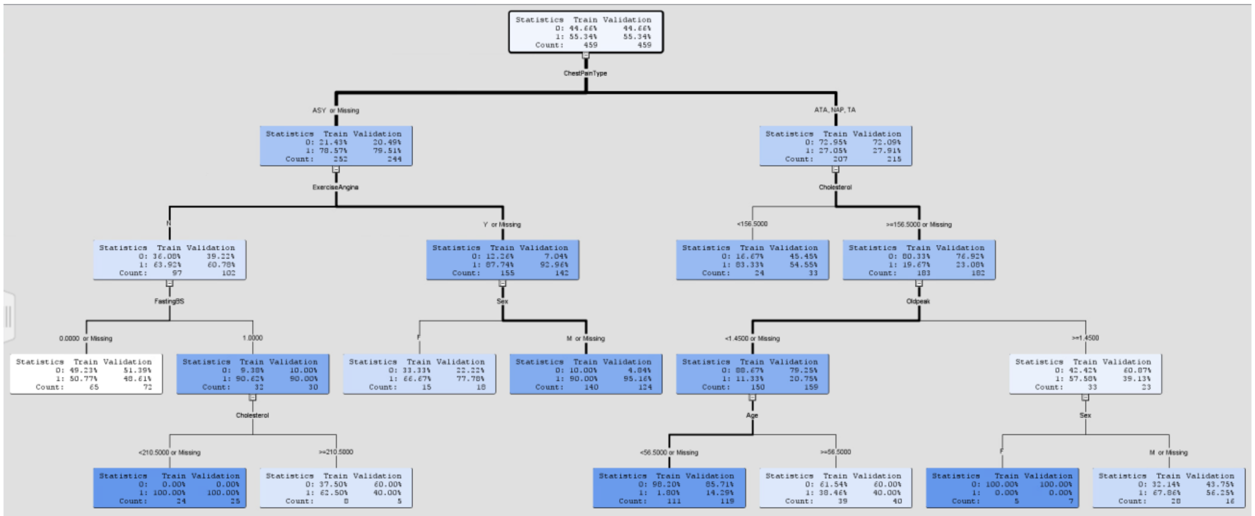
Procedure:

1. Drag the **Decision Tree** node from the Model tab. Drag it into the diagram and connect it to the **Data Partition** node.



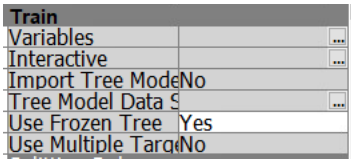
2. Do not make any changes to the properties.
3. Select Interactive in the properties panel and open the decision tree.
4. Right click on the root node and select Train Node. This will create the maximal tree.





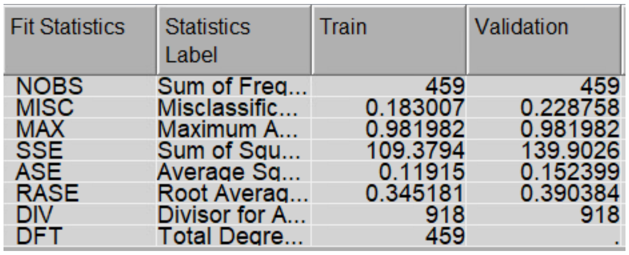
The maximal tree for our data has 10 leaf nodes. The initial split for this decision tree is ChestPainType, dividing between Asymptomatic pain (ASY) with 79% predictiveness of heart disease and Typical Angina (TA), Atypical Angina (ATA), and Non-Anginal Pain (NAP) with only 27% predictiveness. The ASY pain was then divided by Exercise Angina, with “Yes” responses (90% predictiveness) then being split by Sex, with males having higher risk of heart disease.

5. Save the maximal tree and exit Interactive.
6. Freeze the maximal tree in the properties panel.



Train	
Variables	...
Interactive	...
Import Tree Mode	No
Tree Model Data S	...
Use Frozen Tree	Yes
Use Multiple Targ	No

7. Run the node and open Fit Statistics in Results.



Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Freq...	459	459
MISC	Misclassific...	0.183007	0.228758
MAX	Maximum A...	0.981982	0.981982
SSE	Sum of Squ...	109.3794	139.9026
ASE	Average Sq...	0.11915	0.152399
RASE	Root Averag...	0.345181	0.390384
DIV	Divisor for A...	918	918
DFT	Total Degre...	459	.

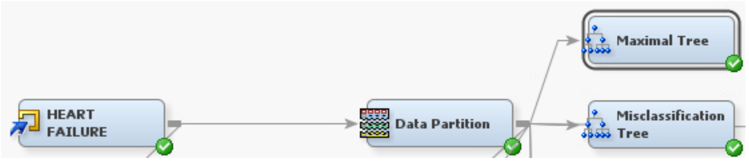
The maximal tree has a Misclassification Rate of 0.228758 and an ASE of 0.152399.

Average Squared Error Tree and Misclassification Tree:

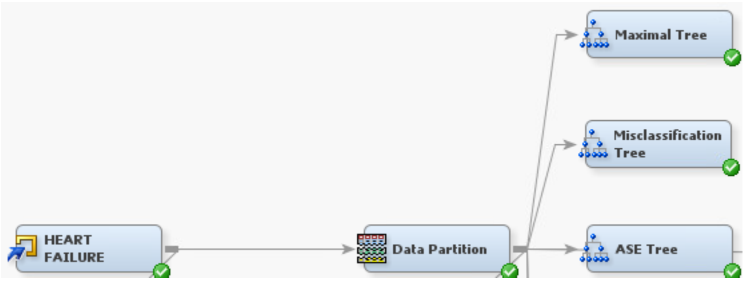
This is an optimal decision tree created by selecting the Average squared error (ASE) and Misclassification rate as assessment measures, respectively.

Procedure:

1. Drag the **Decision Tree** node into the diagram and connect it to the **Data Partition** node.

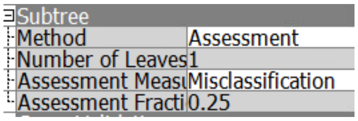


Misclassification

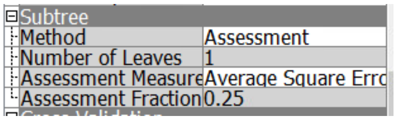


ASE

2. Select the assessment measures to ASE for the Average Squared Error decision tree and Misclassification rate for the Misclassification tree.



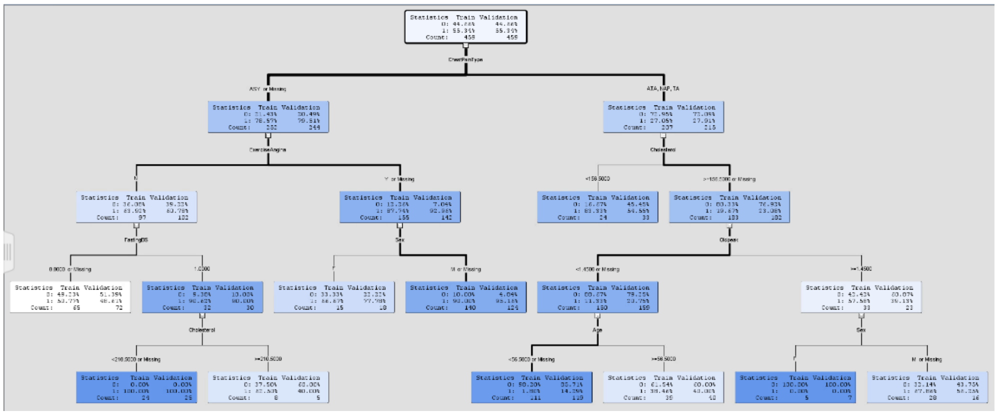
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measu	Misclassification
Assessment Fracti	0.25

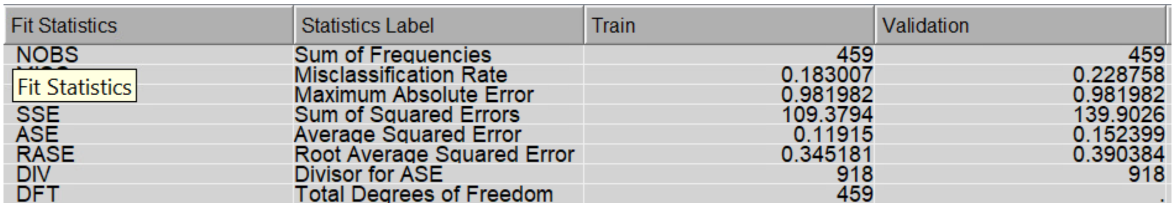


Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Erro
Assessment Fraction	0.25

3. Run the nodes
4. Save both trees then freeze them in the properties panel.

Misclassification tree:

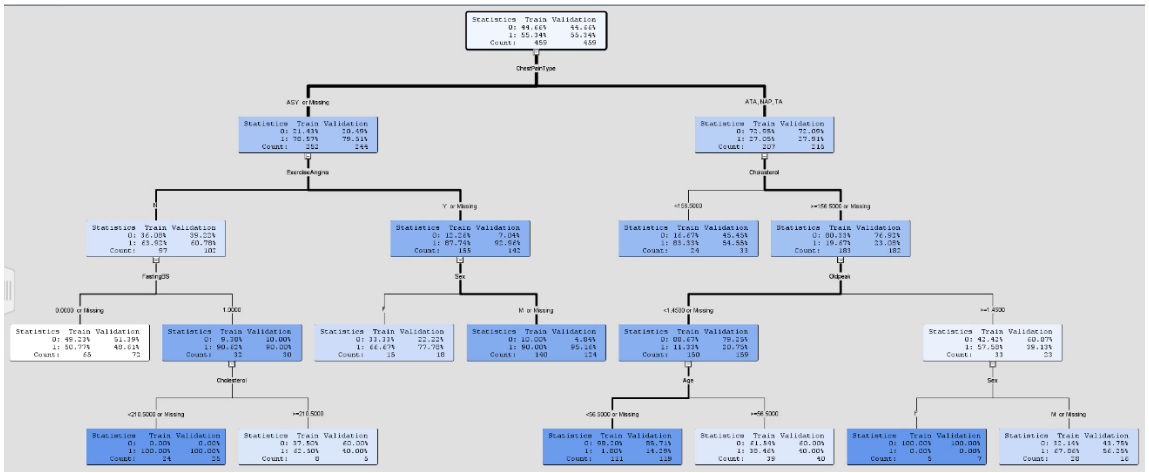


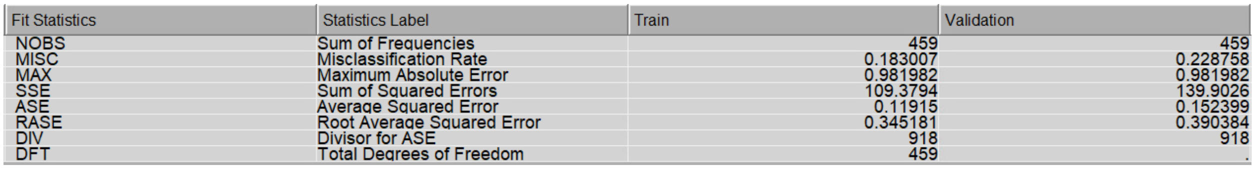


Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	459	459
Fit Statistics	Misclassification Rate	0.183007	0.228758
	Maximum Absolute Error	0.981982	0.981982
SSE	Sum of Squared Errors	109.3794	139.9026
ASE	Average Squared Error	0.11915	0.152399
RASE	Root Average Squared Error	0.345181	0.390384
DIV	Divisor for ASE	918	918
DFT	Total Degrees of Freedom	459	.

Misclassification Tree has a Misclassification Rate of 0.228758 and an ASE of 0.152399.

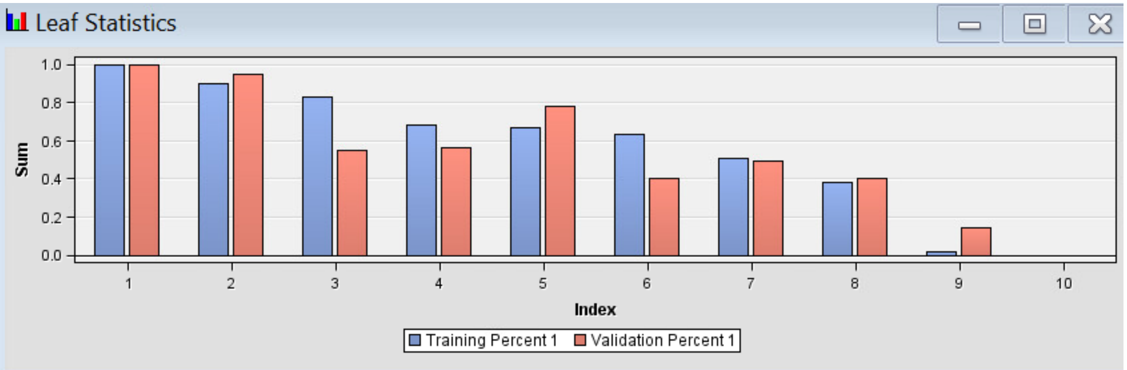
Average Squared Error Tree:





Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	459	459
MISC	Misclassification Rate	0.183007	0.228758
MAX	Maximum Absolute Error	0.981982	0.981982
SSE	Sum of Squared Errors	109.3794	139.9026
ASE	Average Squared Error	0.11915	0.152399
RASE	Root Average Squared Error	0.345181	0.390384
DIV	Divisor for ASE	918	918
DFT	Total Degrees of Freedom	459	.

ASE Tree has a Misclassification Rate of 0.228758 and an ASE of 0.152399.



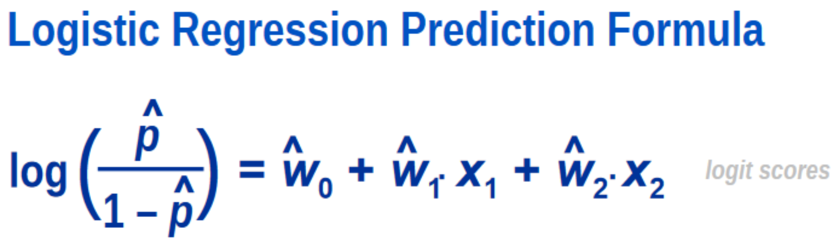
This is the leaf statistics for all three of our decision trees. Statistics remain the same throughout each decision tree.

Conclusion:

On the basis of Average Squared Error and Misclassification Rate, we can conclude that all of our trees have an equal level of predictiveness, as they all have the same Misclassification Rate and ASE. Both the ASE tree and the Misclassification tree have the same division of splits as the Maximal tree as well. Additionally, all of our trees have the same leaf statistics. This is an understandable result, considering we are working with such a small amount of data. What these trees seem to indicate is that the type of chest pain you are experiencing is most indicative of heart disease and that Males are at increase risk than Females.

Regressions:

A linear regression model will be used, only, if our target has an interval variable. However, if our target has a binary value then our model of interest will be logistic regression. The logistic regression model uses the following prediction formula.



Logistic Regression Prediction Formula

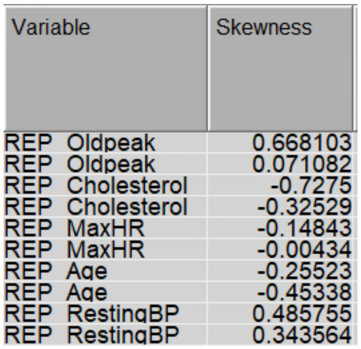
$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 \cdot x_1 + \hat{w}_2 \cdot x_2$$

logit scores

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution.

Skewness and Transformations:

Before we create our regression models, we must check for any extreme outliers in our data that could skew the models and reduce their performance. Any outliers must be transformed logarithmically. To check for skewness, we took a closer look at the Interval Variables of our **Replacement** node.



Variable	Skewness
REP_Oldpeak	0.668103
REP_Oldpeak	0.071082
REP_Cholesterol	-0.7275
REP_Cholesterol	-0.32529
REP_MaxHR	-0.14843
REP_MaxHR	-0.00434
REP_Age	-0.25523
REP_Age	-0.45338
REP_RestingBP	0.485755
REP_RestingBP	0.343564

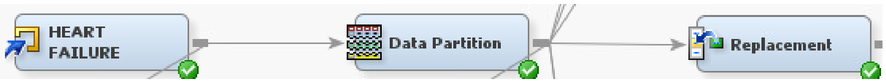
From this table we can see that there are no extreme outliers that would skew our data. Therefore, we do not need to run any transformations and can continue on to our regression models.

Extreme Values:

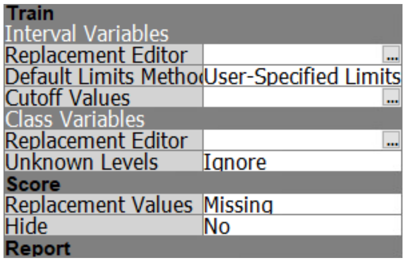
When examining our dataset we noticed some extreme values in Cholesterol, MaxHR, and Resting BP. To control for this we used a **Replacement** node to set a cap and floor for these variable values. Values outside of these cutoffs will be labeled as Missing. We will then impute these missing values to replace them with an estimated value based on the mean of other variable values.

Procedure:

1. Drag the **Replacement** node from the Modify tab into the diagram and connect to the **Data Partition** node.

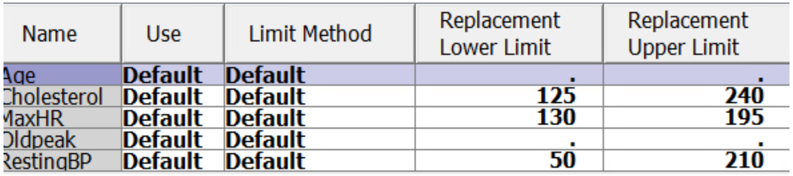


2. Set the Default Limits Method to User Specified Limits and Replacement Value to Missing.



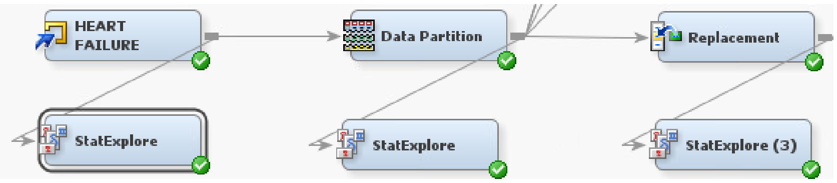
Train	
Interval Variables	
Replacement Editor	...
Default Limits Metho	User-Specified Limits
Cutoff Values	...
Class Variables	
Replacement Editor	...
Unknown Levels	Ignore
Score	
Replacement Values	Missing
Hide	No
Report	

3. Click on the Replacement Editor and make the following changes.

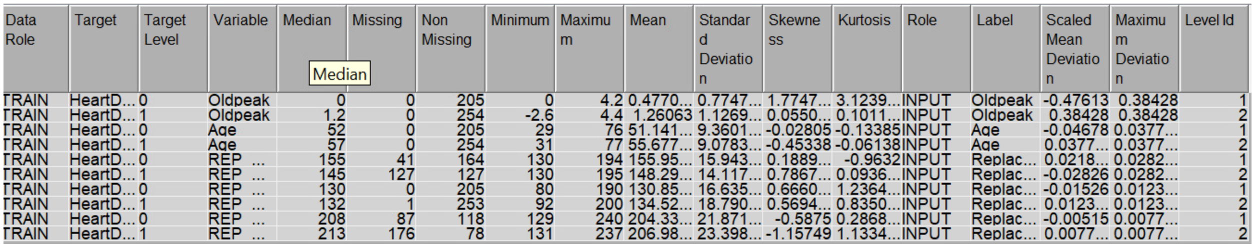


Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit
Age	**Default**	**Default**	**.**	**.**
Cholesterol	**Default**	**Default**	**125**	**240**
MaxHR	**Default**	**Default**	**130**	**195**
Oldpeak	**Default**	**Default**	**.**	**.**
RestingBP	**Default**	**Default**	**50**	**210**

4. Connect a **StatExplore** node from the Sample tab to the **Replacement** node.

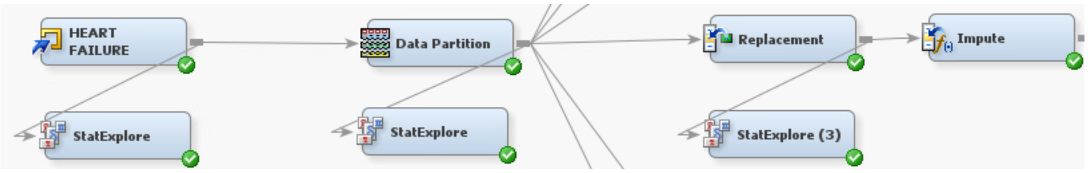


5. Run **StatExplore** and select View > Summary Statistics > Interval Variables.

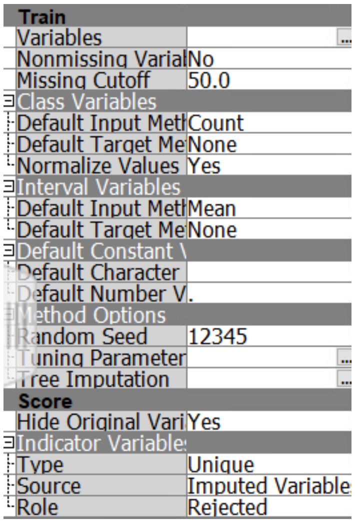


Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	HeartD...	0	Oldpeak	0	0	205	0	4.2	0.4770...	0.7747...	1.7747...	3.1239...	INPUT	Oldpeak	-0.47613	0.38428	1
TRAIN	HeartD...	1	Oldpeak	1.2	0	254	-2.6	4.4	1.26063	1.1269...	0.0550...	0.1011...	INPUT	Oldpeak	0.38428	0.38428	2
TRAIN	HeartD...	0	Age	52	0	205	29	76	51.141...	9.3601...	-0.02805	-0.13385	INPUT	Age	-0.04678	0.0377...	1
TRAIN	HeartD...	1	Age	57	0	254	31	77	55.677...	9.0783...	-0.45338	-0.06138	INPUT	Age	0.0377...	0.0377...	2
TRAIN	HeartD...	0	REP ...	155	41	164	130	194	155.95...	15.943...	0.1889...	-0.9632	INPUT	Replac...	0.0218...	0.0282...	1
TRAIN	HeartD...	1	REP ...	145	127	127	130	195	148.29...	14.117...	0.7867...	0.0936...	INPUT	Replac...	-0.02826	0.0282...	2
TRAIN	HeartD...	0	REP ...	130	0	205	80	190	130.85...	16.635...	0.6660...	1.2364...	INPUT	Replac...	-0.01526	0.0123...	1
TRAIN	HeartD...	1	REP ...	132	1	253	92	200	134.52...	18.790...	0.5694...	0.8350...	INPUT	Replac...	0.0123...	0.0123...	2
TRAIN	HeartD...	0	REP ...	208	87	118	129	240	204.33...	21.871...	-0.5875	0.2868...	INPUT	Replac...	-0.00515	0.0077...	1
TRAIN	HeartD...	1	REP ...	213	176	78	131	237	206.98...	23.398...	-1.15749	1.1334...	INPUT	Replac...	0.0077...	0.0077...	2

6. Drag the **Impute** node from the Modify tab and connect it to the **Replacement** node.



7. Set **Type** to **Unique.** This will ensure that missing values are replaced with unique values that will allow us to determine if these missing values are important for our analysis.



Train	
Variables	...
Nonmissing Varial	No
Missing Cutoff	50.0
Class Variables	
Default Input Met	Count
Default Target Me	None
Normalize Values	Yes
Interval Variables	
Default Input Met	Mean
Default Target Me	None
Default Constant \	
Default Character	
Default Number V	.
Method Options	
Random Seed	12345
Tuning Parameter	...
Tree Imputation	...
Score	
Hide Original Vari	Yes
Indicator Variable	
Type	Unique
Source	Imputed Variable
Role	Rejected

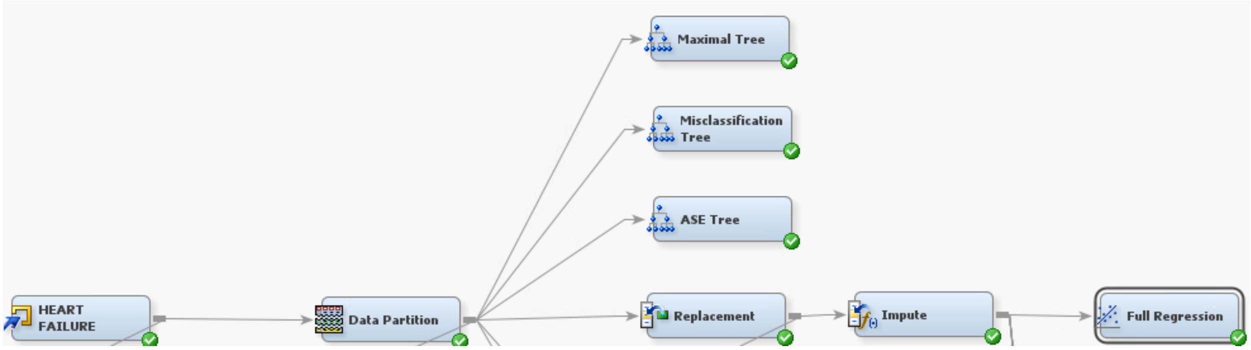
Regression Models:

We have done the following four types of regression in this project:

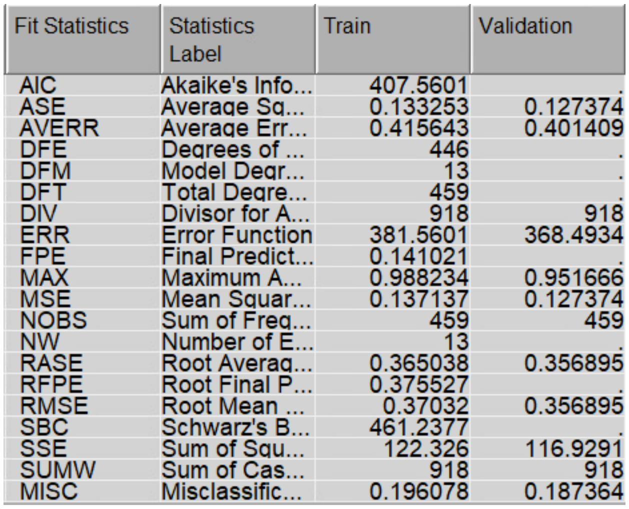
- Full Regression
- Forward Inclusion
- Backward Exclusion
- Stepwise Regression

Full Regression:

1. Drag the **Regression** node and connect it to the **Impute** node.



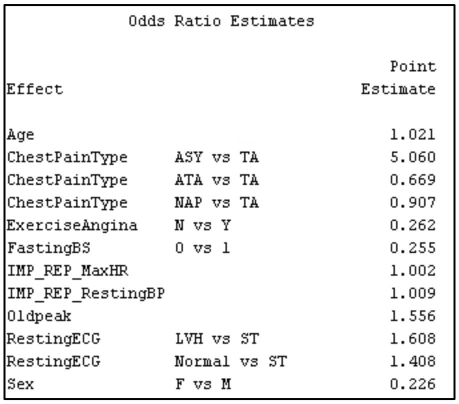
2. Run the node.



Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Info...	407.5601	.
ASE	Average Sq...	0.133253	0.127374
AVERR	Average Err...	0.415643	0.401409
DFE	Degrees of ...	446	.
DFM	Model Degr...	13	.
DFT	Total Degre...	459	.
DIV	Divisor for A...	918	918
ERR	Error Function	381.5601	368.4934
FPE	Final Predict...	0.141021	.
MAX	Maximum A...	0.988234	0.951666
MSE	Mean Squar...	0.137137	0.127374
NOBS	Sum of Freq...	459	459
NW	Number of E...	13	.
RASE	Root Averag...	0.365038	0.356895
RFPE	Root Final P...	0.375527	.
RMSE	Root Mean ...	0.37032	0.356895
SBC	Schwarz's B...	461.2377	.
SSE	Sum of Squ...	122.326	116.9291
SUMW	Sum of Cas...	918	918
MISC	Misclassific...	0.196078	0.187364

ASE is 0.127374.

3. Open the Output and scroll to Odds Ration Estimate.



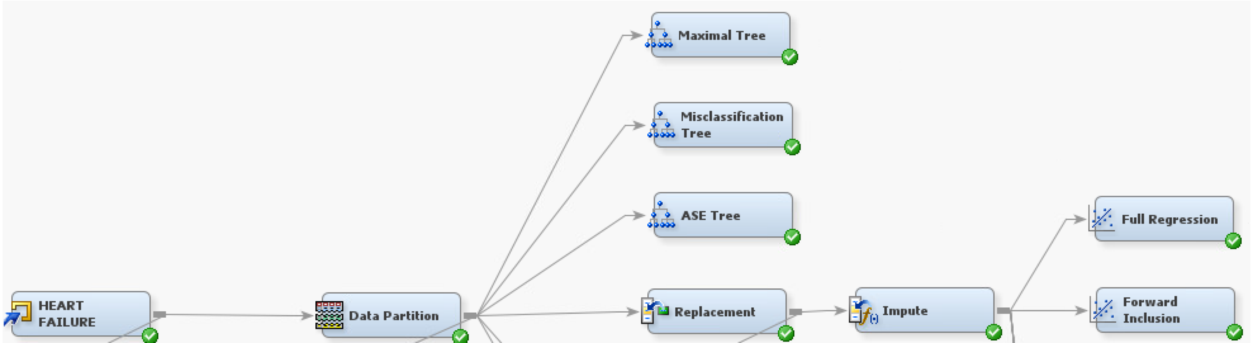
```
              Odds Ratio Estimates

                                            Point
Effect                                   Estimate

Age                                         1.021
ChestPainType      ASY vs TA                5.060
ChestPainType      ATA vs TA                0.669
ChestPainType      NAP vs TA                0.907
ExerciseAngina     N vs Y                   0.262
FastingBS          0 vs 1                   0.255
IMP_REP_MaxHR                               1.002
IMP_REP_RestingBP                           1.009
Oldpeak                                     1.556
RestingECG         LVH vs ST                1.608
RestingECG         Normal vs ST             1.408
Sex                F vs M                   0.226
```

ASY ChestPainType is 5.06 times more indicative of heart disease than TA ChestPainType. For Resting ECG, Left Ventricular Hypertrophy (LVH) is 1.608 times more indicative of heart disease than ST segmentation (ST).

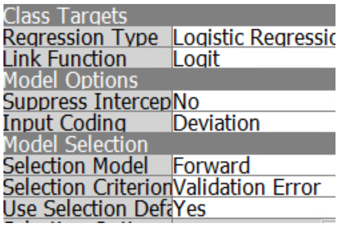
Procedure for Forward Inclusion, Backward Exclusion, Stepwise Regression:

For all these regressions models we will repeat step one of the Full Regression procedure. However, in step two we will be proceeding as follows:

Forward Inclusion:

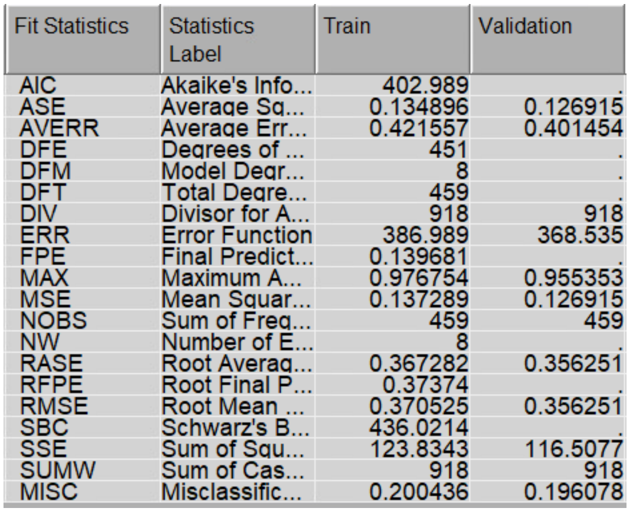


2. In the properties panel, select model as Forward and set Selection Criterion to Validation Error.



Class Targets	
Regression Type	Logistic Regressic
Link Function	Logit
Model Options	
Suppress Intercep	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defa	Yes

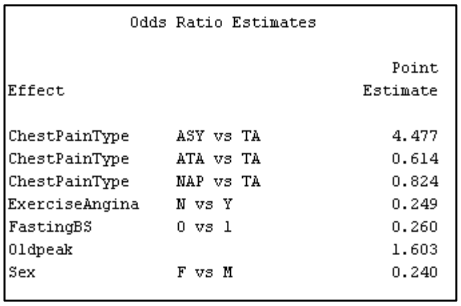
3. Run the node.



Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Info...	402.989	.
ASE	Average Sq...	0.134896	0.126915
AVERR	Average Err...	0.421557	0.401454
DFE	Degrees of ...	451	.
DFM	Model Degr...	8	.
DFT	Total Degre...	459	.
DIV	Divisor for A...	918	918
ERR	Error Function	386.989	368.535
FPE	Final Predict...	0.139681	.
MAX	Maximum A...	0.976754	0.955353
MSE	Mean Squar...	0.137289	0.126915
NOBS	Sum of Freq...	459	459
NW	Number of E...	8	.
RASE	Root Averag...	0.367282	0.356251
RFPE	Root Final P...	0.37374	.
RMSE	Root Mean ...	0.370525	0.356251
SBC	Schwarz's B...	436.0214	.
SSE	Sum of Squ...	123.8343	116.5077
SUMW	Sum of Cas...	918	918
MISC	Misclassific...	0.200436	0.196078

ASE is 0.126915, better than the Full Regression model.

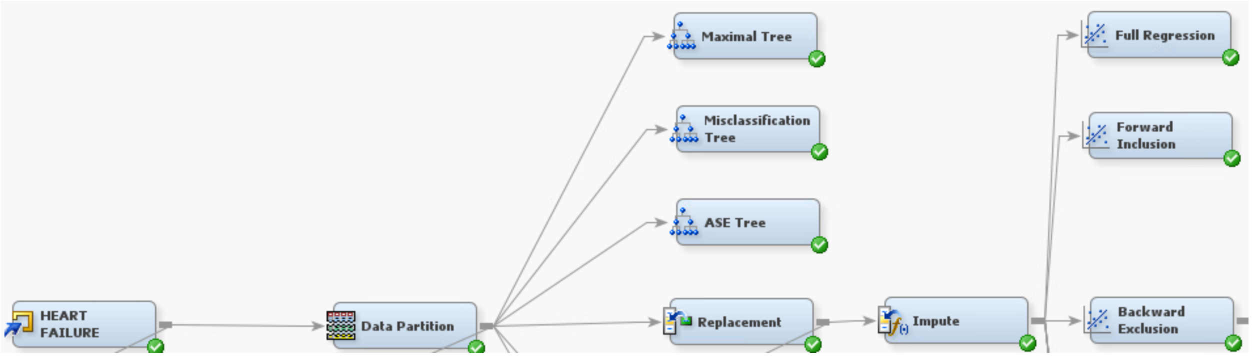
4. Open the Output and scroll to Odds Ratio Estimate.



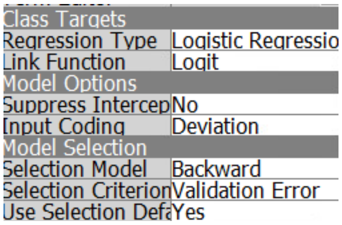
```
             Odds Ratio Estimates

                                          Point
Effect                                 Estimate

ChestPainType      ASY vs TA              4.477
ChestPainType      ATA vs TA              0.614
ChestPainType      NAP vs TA              0.824
ExerciseAngina     N vs Y                 0.249
FastingBS          0 vs 1                 0.260
Oldpeak                                   1.603
Sex                F vs M                 0.240
```

ASY ChestPainType is 4.477 times more indicative of heart disease than TA ChestPainType. TA ChestPainType is 1.63x (1/0.614) more indicative of heart disease than ATA ChestPainType.

Backward Exclusion:

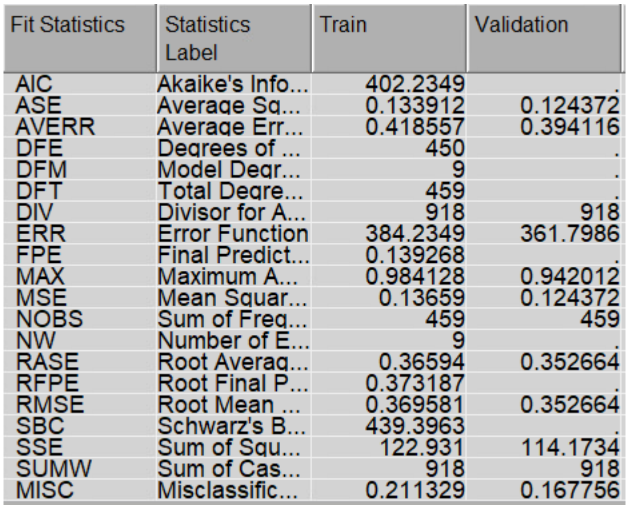


2. In the model selection properties select model as Backward and Selection Criterion as Validation Error as shown in the picture below.



Class Targets	
Regression Type	Logistic Regressio
Link Function	Logit
Model Options	
Suppress Intercep	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Error
Use Selection Defa	Yes

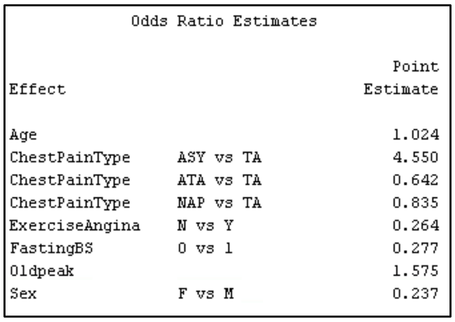
3. Run the node.



Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Info...	402.2349	.
ASE	Average Sq...	0.133912	0.124372
AVERR	Average Err...	0.418557	0.394116
DFE	Degrees of ...	450	.
DFM	Model Degr...	9	.
DFT	Total Degre...	459	.
DIV	Divisor for A...	918	918
ERR	Error Function	384.2349	361.7986
FPE	Final Predict...	0.139268	.
MAX	Maximum A...	0.984128	0.942012
MSE	Mean Squar...	0.13659	0.124372
NOBS	Sum of Freq...	459	459
NW	Number of E...	9	.
RASE	Root Averag...	0.36594	0.352664
RFPE	Root Final P...	0.373187	.
RMSE	Root Mean ...	0.369581	0.352664
SBC	Schwarz's B...	439.3963	.
SSE	Sum of Squ...	122.931	114.1734
SUMW	Sum of Cas...	918	918
MISC	Misclassific...	0.211329	0.167756

ASE is 0.124372. This is a better model than Full Regression and Forward Regression.

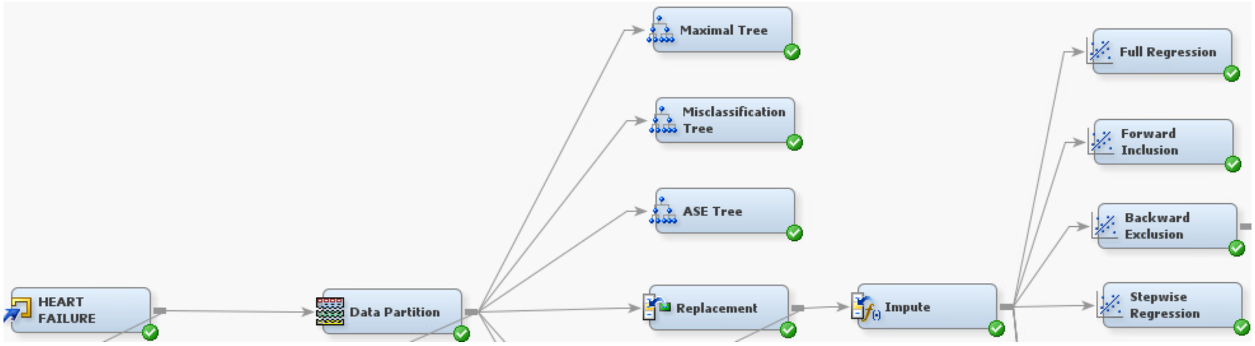
4. Open the Output and scroll to Odds Ratio Estimate.



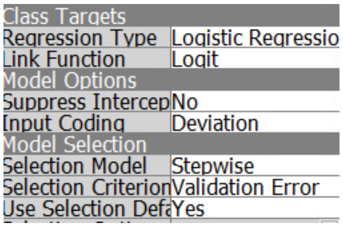
```
              Odds Ratio Estimates

                                           Point
Effect                                  Estimate

Age                                        1.024
ChestPainType      ASY vs TA               4.550
ChestPainType      ATA vs TA               0.642
ChestPainType      NAP vs TA               0.835
ExerciseAngina     N vs Y                  0.264
FastingBS          0 vs 1                  0.277
Oldpeak                                    1.575
Sex                F vs M                  0.237
```

ASY ChestPainType is 4.550 times more indicative of heart disease than TA ChestPainType. ExerciseAngina indicates that you are 3.8x (1/0.264) more likely to have heart disease. Females are less prone to have heart disease than Males by 76%.

Stepwise Regression:

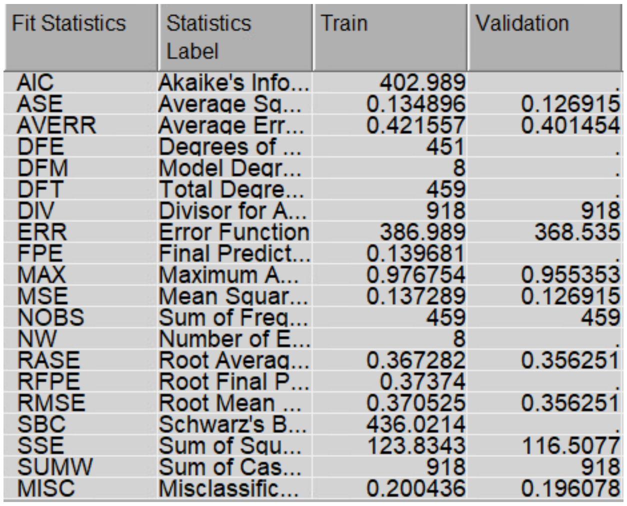


2. In the model selection properties select model as Stepwise and Selection Criterion as Validation Error.



Class Targets	
Regression Type	Logistic Regressio
Link Function	Logit
Model Options	
Suppress Intercep	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defa	Yes

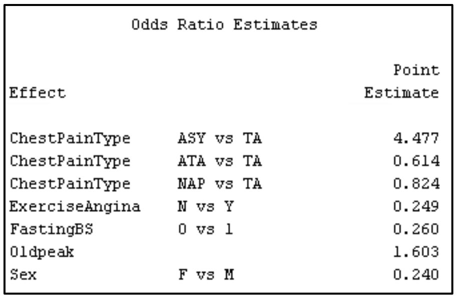
3. Run the node.



Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Info...	402.989	.
ASE	Average Sq...	0.134896	0.126915
AVERR	Average Err...	0.421557	0.401454
DFE	Degrees of ...	451	.
DFM	Model Degr...	8	.
DFT	Total Degre...	459	.
DIV	Divisor for A...	918	918
ERR	Error Function	386.989	368.535
FPE	Final Predict...	0.139681	.
MAX	Maximum A...	0.976754	0.955353
MSE	Mean Squar...	0.137289	0.126915
NOBS	Sum of Freq...	459	459
NW	Number of E...	8	.
RASE	Root Averag...	0.367282	0.356251
RFPE	Root Final P...	0.37374	.
RMSE	Root Mean ...	0.370525	0.356251
SBC	Schwarz's B...	436.0214	.
SSE	Sum of Squ...	123.8343	116.5077
SUMW	Sum of Cas...	918	918
MISC	Misclassific...	0.200436	0.196078

ASE is 0.126915, not as good as the Backwards Regression model.

4. Open the Output and scroll to Odds Ratio Estimate.



```
             Odds Ratio Estimates

                                          Point
Effect                                 Estimate

ChestPainType     ASY vs TA               4.477
ChestPainType     ATA vs TA               0.614
ChestPainType     NAP vs TA               0.824
ExerciseAngina    N vs Y                  0.249
FastingBS         0 vs 1                  0.260
Oldpeak                                   1.603
Sex               F vs M                  0.240
```

ASY ChestPainType is 4.477 times more indicative of heart disease than TA ChestPainType. For every elevation in Oldpeak levels, you are 1.603x more likely to experience heart disease. Males are more prone to heart disease than females.

Conclusion:

Based on average squared error (ASE) we can conclude that Backward Exclusion is the best regression model among all the regression models as it has the lowest average squared error among all the regression models. Therefore, this is the model we will be using to optimize two of our neural network models. Asymptomatic ChestPainType is highly predictive of heart disease, as well as gender and Oldpeak levels which indicate previous heart attacks.

Neural Networks:

A neural network is a set of connected input/output variables where each connection has a given weight that determines the outcome. Neural networks take non-linear functions of linear combinations of input variables. This is a powerful and very general approach for regression and classification and has been shown to be the best machine learning method on many problems.

Problems in Neural Networks:

- Extreme or unusual values also present a problem for neural networks. The problem is mitigated somewhat by the hyperbolic tangent activation functions in the hidden units
- It cannot select its input; however, this can be reduced by the complexity optimization algorithm called “stopped running” which minimized the chance of overfitting.
- It is not possible to interpret the input variable of Neural Network.

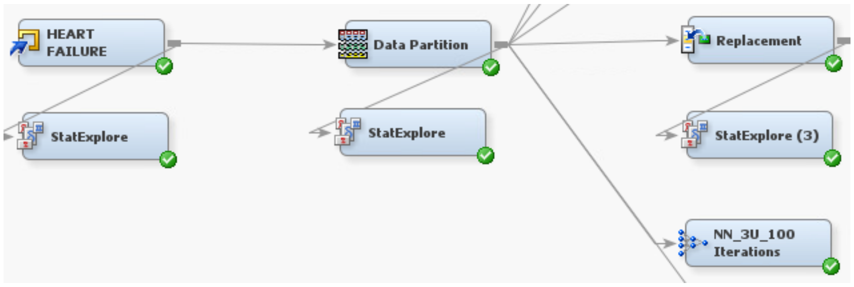
Neural Network Models:

- Neural Network with 3 hidden units and 100 iterations (with and without replacement)
- Neural Network with 3 hidden units and 100 iterations (with and without replacement)
- Neural network with 4 hidden units and 50 iterations
- Neural Network with 3 hidden units and preliminary training
- 3 Hidden unit Neural Network connected to Backwards Regression (100 iterations)
- 3 Hidden unit Neural Network connected to Backwards Regression (50 iterations)

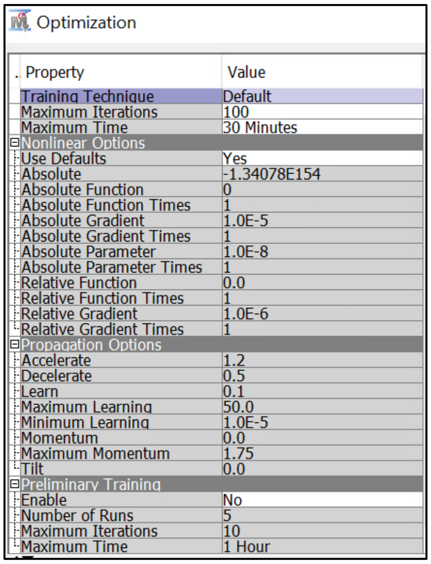
3 Hidden Unit Neural Network (100 iterations, no Replacement):

Procedure:

1. Drag a **Neural Network** node into the diagram from the Model tab and connect to the **Data Partition.**

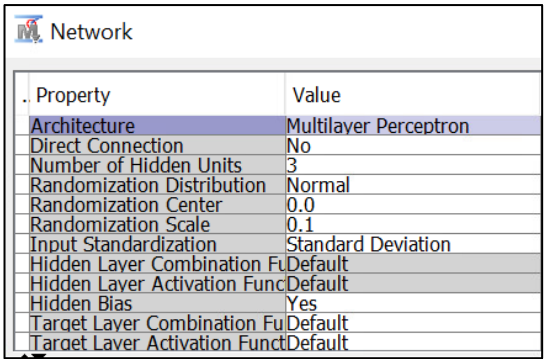


2. Select **Optimization**, from the properties panel and disable preliminary training. Set maximum iterations to 100. Close Optimization.


Optimization

Property	Value
Training Technique	Default
Maximum Iterations	100
Maximum Time	30 Minutes
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1
Relative Function	0.0
Relative Function Times	1
Relative Gradient	1.0E-6
Relative Gradient Times	1
Propagation Options	
Accelerate	1.2
Decelerate	0.5
Learn	0.1
Maximum Learning	50.0
Minimum Learning	1.0E-5
Momentum	0.0
Maximum Momentum	1.75
Tilt	0.0
Preliminary Training	
Enable	No
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

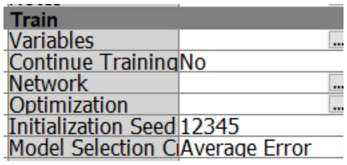
3. Select **Network** from the properties panel and set hidden units to 3.



Network

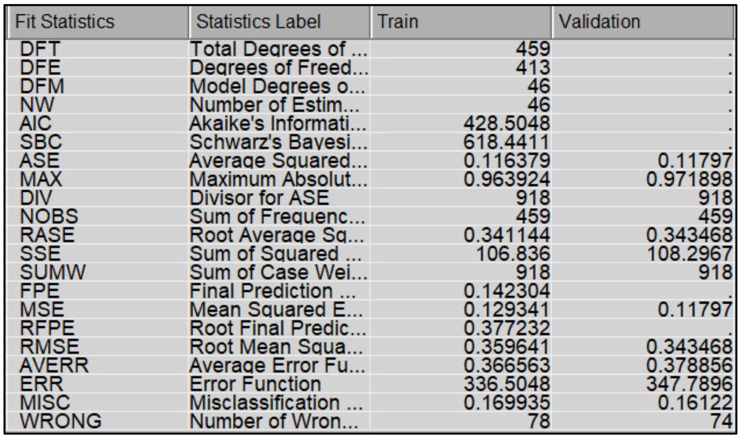
Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	3
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Fu	Default
Hidden Layer Activation Func	Default
Hidden Bias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

4. Select Average Error as the Model Selection Criterion.



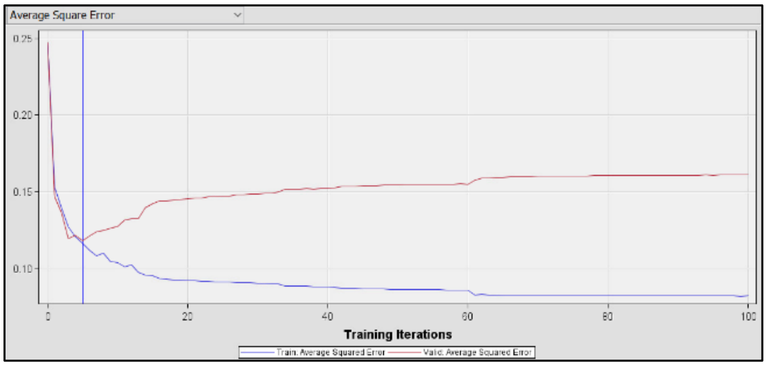
Train	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection C	Average Error

5. Run the Neural Network node and view the results.



Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of ...	459	.
DFE	Degrees of Freed...	413	.
DFM	Model Degrees o...	46	.
NW	Number of Estim...	46	.
AIC	Akaike's Informati...	428.5048	.
SBC	Schwarz's Bayesi...	618.4411	.
ASE	Average Squared...	0.116379	0.11797
MAX	Maximum Absolut...	0.963924	0.971898
DIV	Divisor for ASE	918	918
NOBS	Sum of Frequenc...	459	459
RASE	Root Average Sq...	0.341144	0.343468
SSE	Sum of Squared ...	106.836	108.2967
SUMW	Sum of Case Wei...	918	918
FPE	Final Prediction ...	0.142304	.
MSE	Mean Squared E...	0.129341	0.11797
RFPE	Root Final Predic...	0.377232	.
RMSE	Root Mean Squa...	0.359641	0.343468
AVERR	Average Error Fu...	0.366563	0.378856
ERR	Error Function	336.5048	347.7896
MISC	Misclassification ...	0.169935	0.16122
WRONG	Number of Wron...	78	74

The ASE for this model is 0.11797.



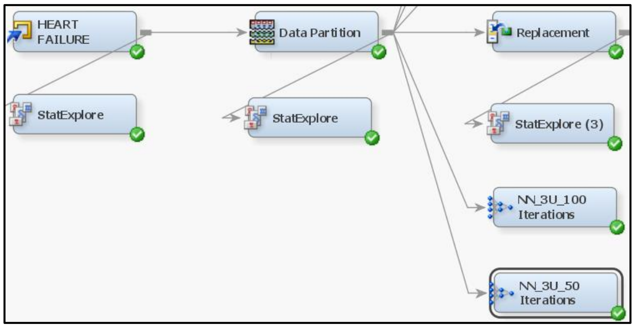
This model converges at 5 iterations.

Note: For the rest of our models, Model Selection Criterion will always be set to Average Error.

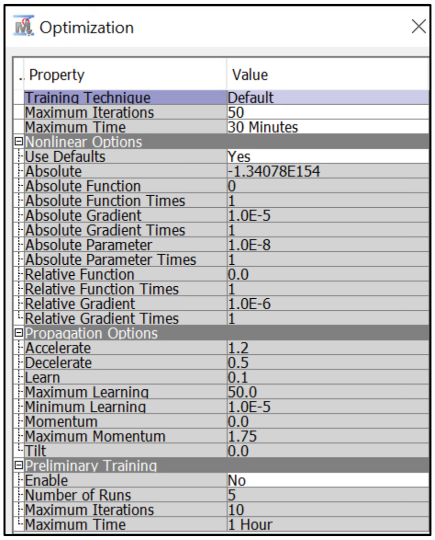
3 Hidden Unit Neural Network (50 iterations, no Replacement):

Procedure:

1. Repeat step one of the 100 Iteration network.

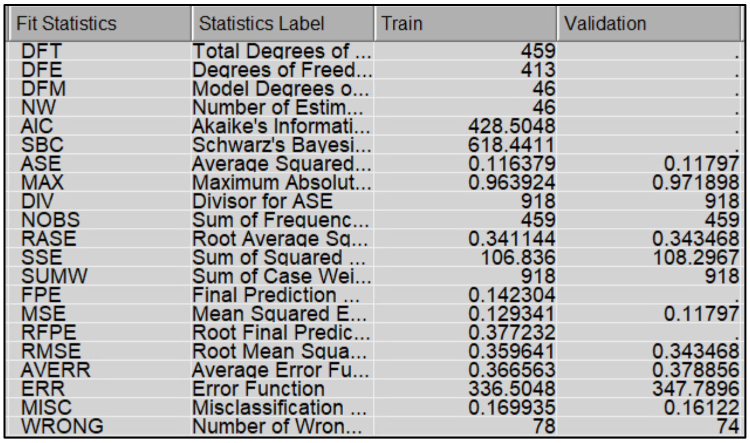


2. Select **Optimization**, from the properties panel and set maximum iterations to 50.


Optimization

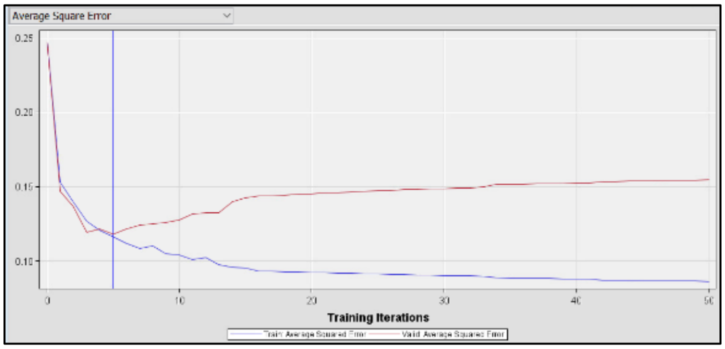
Property	Value
Training Technique	Default
Maximum Iterations	50
Maximum Time	30 Minutes
⊟Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1
Relative Function	0.0
Relative Function Times	1
Relative Gradient	1.0E-6
Relative Gradient Times	1
⊟Propagation Options	
Accelerate	1.2
Decelerate	0.5
Learn	0.1
Maximum Learning	50.0
Minimum Learning	1.0E-5
Momentum	0.0
Maximum Momentum	1.75
Tilt	0.0
⊟Preliminary Training	
Enable	No
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

3. Run the Neural Network node and view the results:



Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of ...	459	.
DFE	Degrees of Freed...	413	.
DFM	Model Degrees o...	46	.
NW	Number of Estim...	46	.
AIC	Akaike's Informati...	428.5048	.
SBC	Schwarz's Bayesi...	618.4411	.
ASE	Average Squared...	0.116379	0.11797
MAX	Maximum Absolut...	0.963924	0.971898
DIV	Divisor for ASE	918	918
NOBS	Sum of Frequenc...	459	459
RASE	Root Average Sq...	0.341144	0.343468
SSE	Sum of Squared ...	106.836	108.2967
SUMW	Sum of Case Wei...	918	918
FPE	Final Prediction ...	0.142304	.
MSE	Mean Squared E...	0.129341	0.11797
RFPE	Root Final Predic...	0.377232	.
RMSE	Root Mean Squa...	0.359641	0.343468
AVERR	Average Error Fu...	0.366563	0.378856
ERR	Error Function	336.5048	347.7896
MISC	Misclassification ...	0.169935	0.16122
WRONG	Number of Wron...	78	74

The ASE for this model is the same as the previous 100 iteration model.

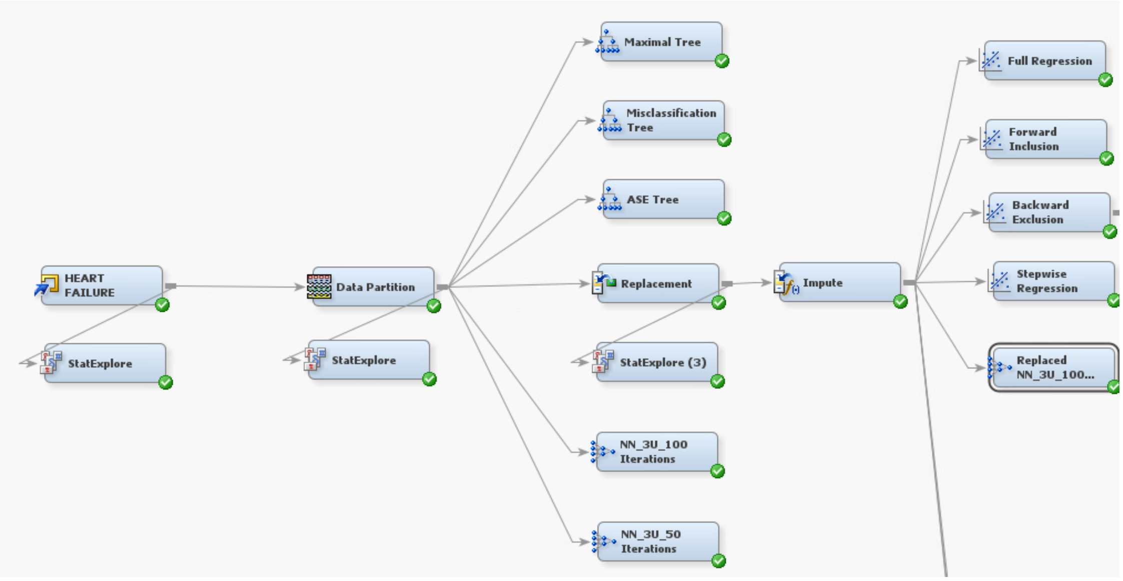


This model converges at 5 iterations.

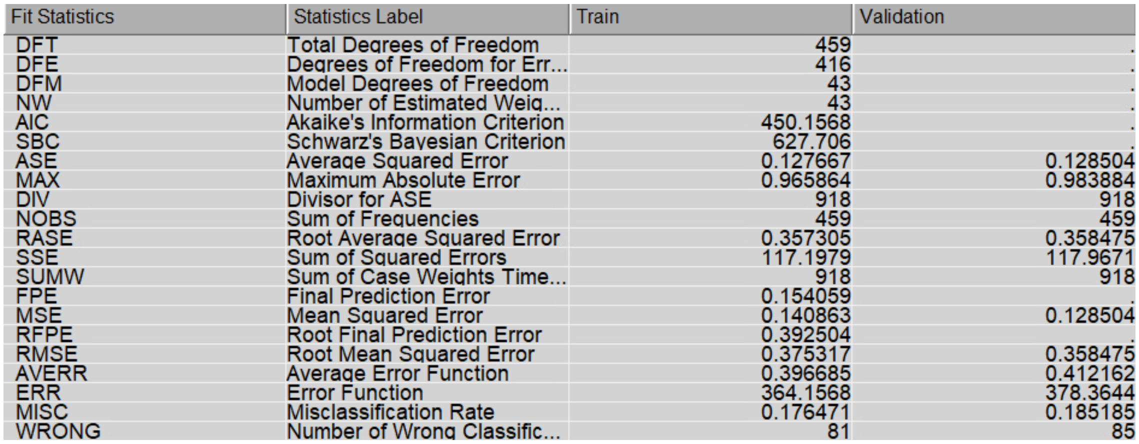
Replaced Neural Network (3 hidden units, 100 iterations):

Procedure:

1. Drag a **Neural Network** node into the diagram and connect it to the **Impute** node.

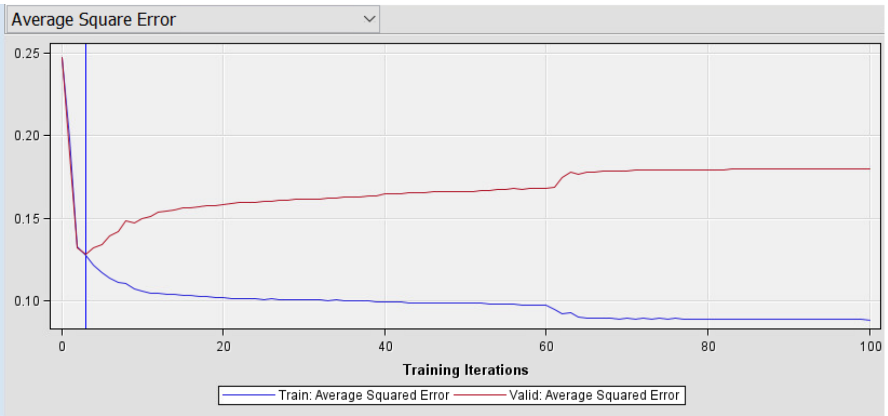


2. Following the same procedure, select 3 hidden units and 100 iterations. Run the results:



Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of Freedom	459	.
DFE	Degrees of Freedom for Err...	416	.
DFM	Model Degrees of Freedom	43	.
NW	Number of Estimated Weig...	43	.
AIC	Akaike's Information Criterion	450.1568	.
SBC	Schwarz's Bayesian Criterion	627.706	.
ASE	Average Squared Error	0.127667	0.128504
MAX	Maximum Absolute Error	0.965864	0.983884
DIV	Divisor for ASE	918	918
NOBS	Sum of Frequencies	459	459
RASE	Root Average Squared Error	0.357305	0.358475
SSE	Sum of Squared Errors	117.1979	117.9671
SUMW	Sum of Case Weights Time...	918	918
FPE	Final Prediction Error	0.154059	.
MSE	Mean Squared Error	0.140863	0.128504
RFPE	Root Final Prediction Error	0.392504	.
RMSE	Root Mean Squared Error	0.375317	0.358475
AVERR	Average Error Function	0.396685	0.412162
ERR	Error Function	364.1568	378.3644
MISC	Misclassification Rate	0.176471	0.185185
WRONG	Number of Wrong Classific...	81	85

The ASE for this model is 0.128504, which is greater than the ASEs for the previous two models.

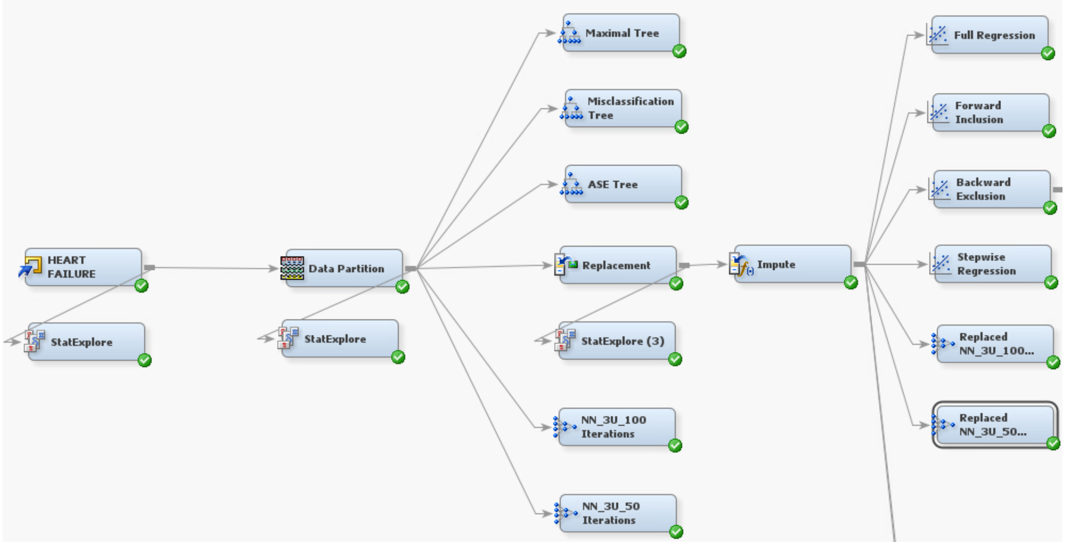


This model converges at 4 iterations.

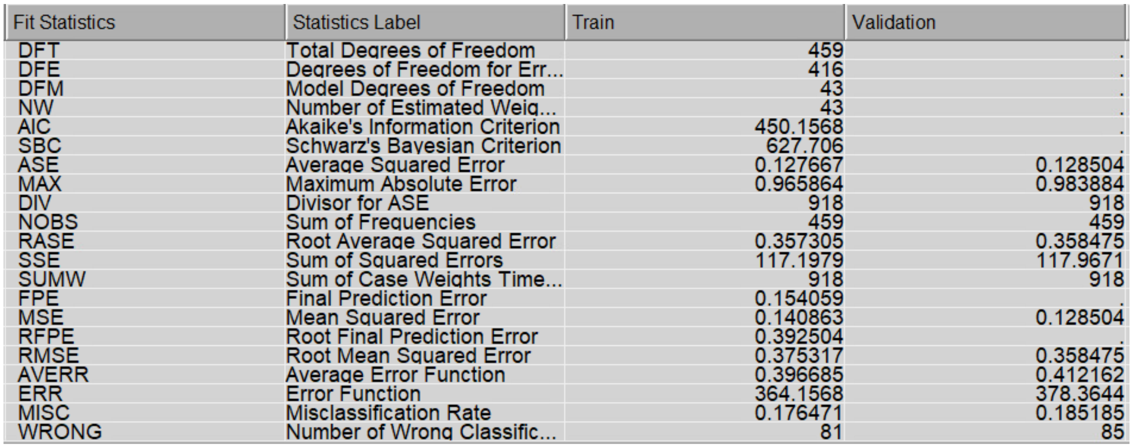
Replaced Neural Network (3 hidden units, 50 iterations):

Procedure:

1. Drag a **Neural Network** node into the diagram and connect it to the **Impute** node:

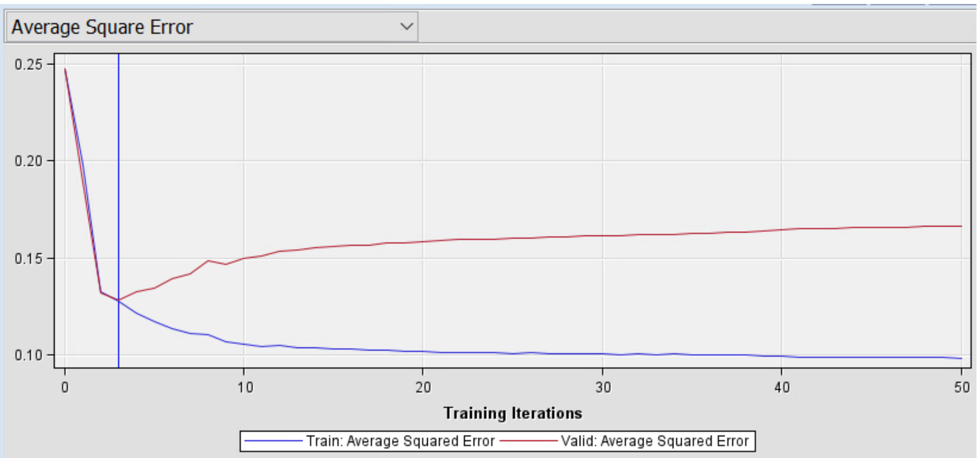


2. Select 3 hidden units and 50 iterations. Run the results:



Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of Freedom	459	.
DFE	Degrees of Freedom for Err...	416	.
DFM	Model Degrees of Freedom	43	.
NW	Number of Estimated Weig...	43	.
AIC	Akaike's Information Criterion	450.1568	.
SBC	Schwarz's Bayesian Criterion	627.706	.
ASE	Average Squared Error	0.127667	0.128504
MAX	Maximum Absolute Error	0.965864	0.983884
DIV	Divisor for ASE	918	918
NOBS	Sum of Frequencies	459	459
RASE	Root Average Squared Error	0.357305	0.358475
SSE	Sum of Squared Errors	117.1979	117.9671
SUMW	Sum of Case Weights Time...	918	918
FPE	Final Prediction Error	0.154059	.
MSE	Mean Squared Error	0.140863	0.128504
RFPE	Root Final Prediction Error	0.392504	.
RMSE	Root Mean Squared Error	0.375317	0.358475
AVERR	Average Error Function	0.396685	0.412162
ERR	Error Function	364.1568	378.3644
MISC	Misclassification Rate	0.176471	0.185185
WRONG	Number of Wrong Classific...	81	85

The ASE for this model is the same as the previous model of 100 iterations.

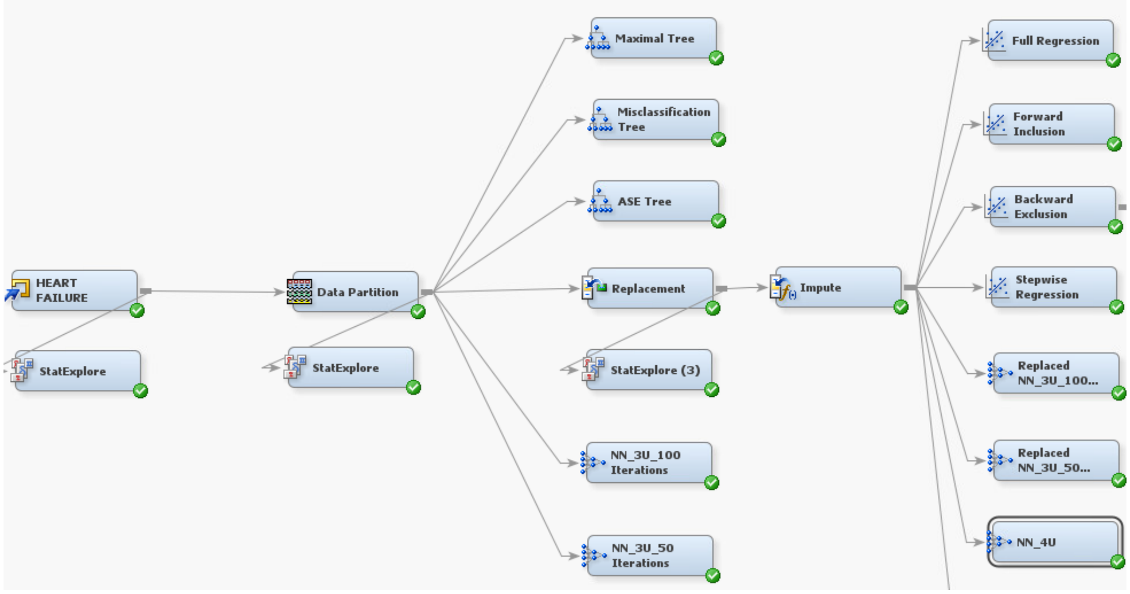


This model converges at 4 iterations.

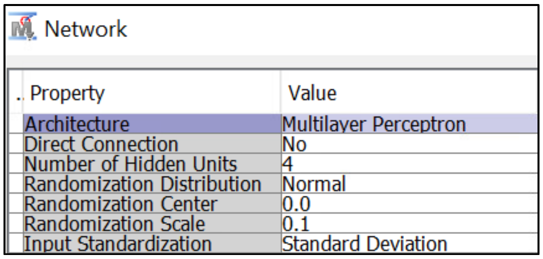
Replaced Neural Network (4 hidden units, 50 iterations):

Procedure:

1. Drag a **Neural Network** tool into the diagram and connect it to the **Impute** node.



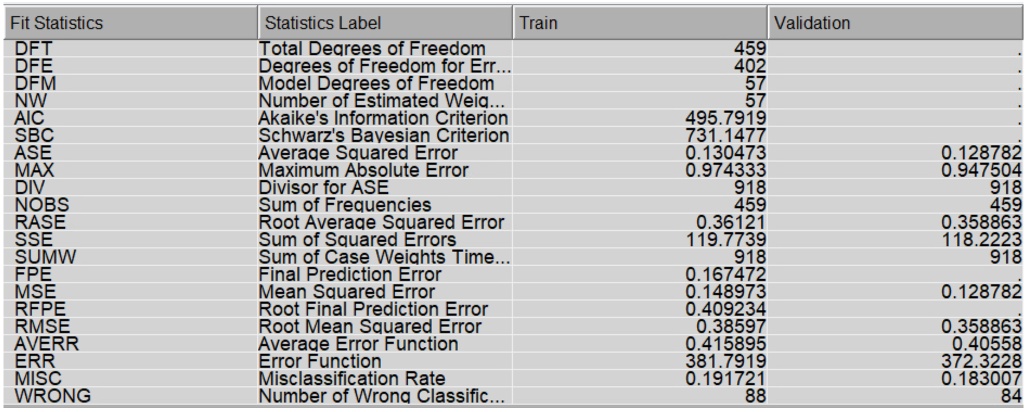
2. Go into the network and set hidden units to 4.



Network

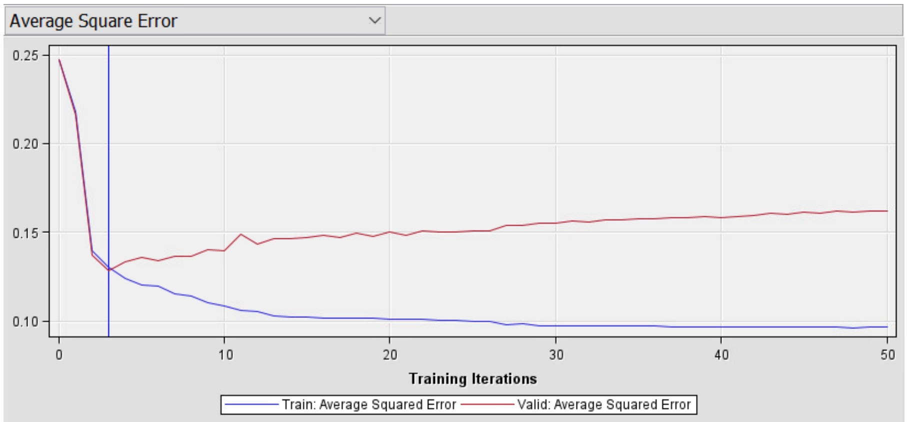
Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	4
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation

3. Run the node:



Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of Freedom	459	.
DFE	Degrees of Freedom for Err...	402	.
DFM	Model Degrees of Freedom	57	.
NW	Number of Estimated Weig...	57	.
AIC	Akaike's Information Criterion	495.7919	.
SBC	Schwarz's Bayesian Criterion	731.1477	.
ASE	Average Squared Error	0.130473	0.128782
MAX	Maximum Absolute Error	0.974333	0.947504
DIV	Divisor for ASE	918	918
NOBS	Sum of Frequencies	459	459
RASE	Root Average Squared Error	0.36121	0.358863
SSE	Sum of Squared Errors	119.7739	118.2223
SUMW	Sum of Case Weights Time...	918	918
FPE	Final Prediction Error	0.167472	.
MSE	Mean Squared Error	0.148973	0.128782
RFPE	Root Final Prediction Error	0.409234	.
RMSE	Root Mean Squared Error	0.38597	0.358863
AVERR	Average Error Function	0.415895	0.40558
ERR	Error Function	381.7919	372.3228
MISC	Misclassification Rate	0.191721	0.183007
WRONG	Number of Wrong Classific...	88	84

The ASE for the 4 hidden unit model is 0.128782. This is larger than the ASE for the 3 hidden unit neural network models. Therefore, we will not be running any models with more than 4 hidden units, as the ASE is only likely to become greater with more hidden units.

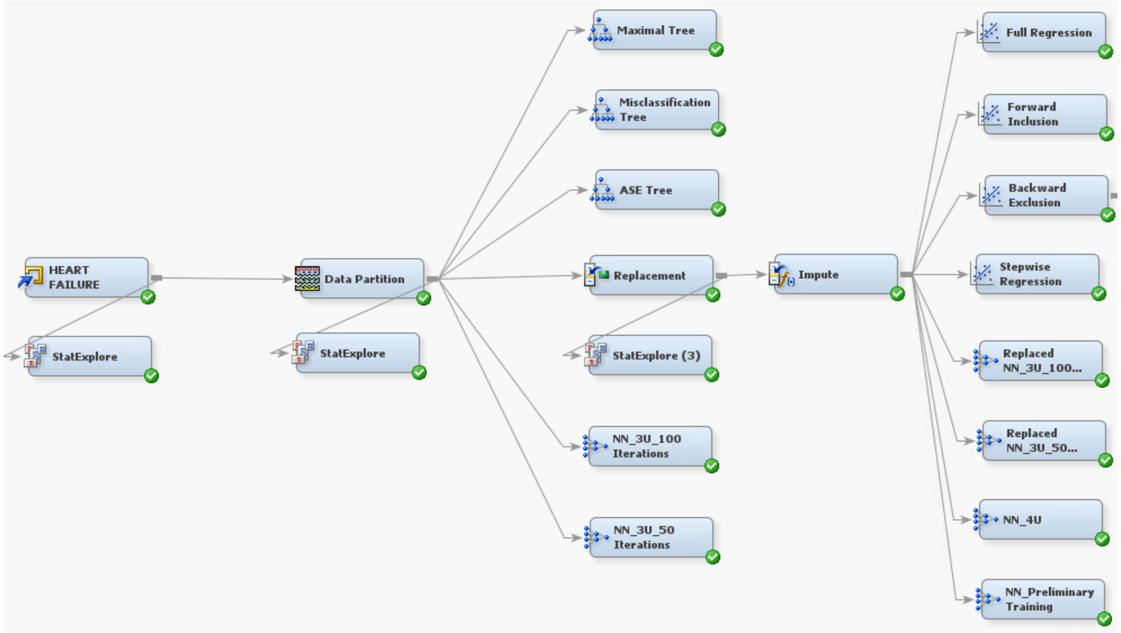


This model converges at 3 iterations.

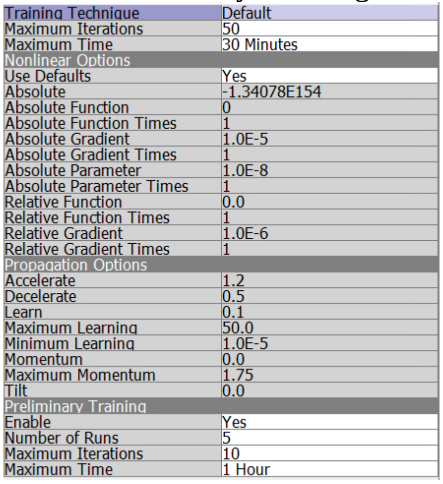
Neural Network with Preliminary Training:

Procedure:

1. Drag a **Neural Network** node into the diagram and connect it to the **Impute** node:

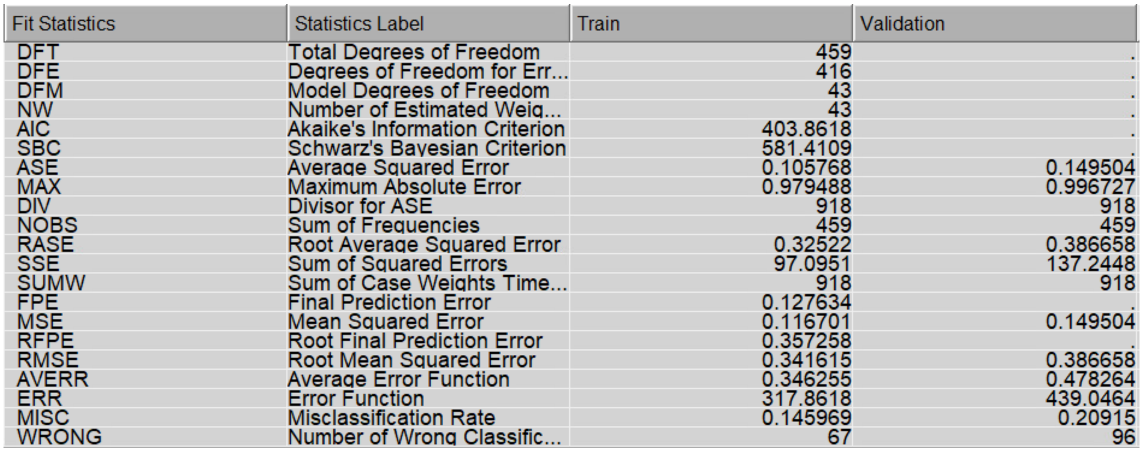


2. Enable Preliminary Training. Set number of iterations to 50.



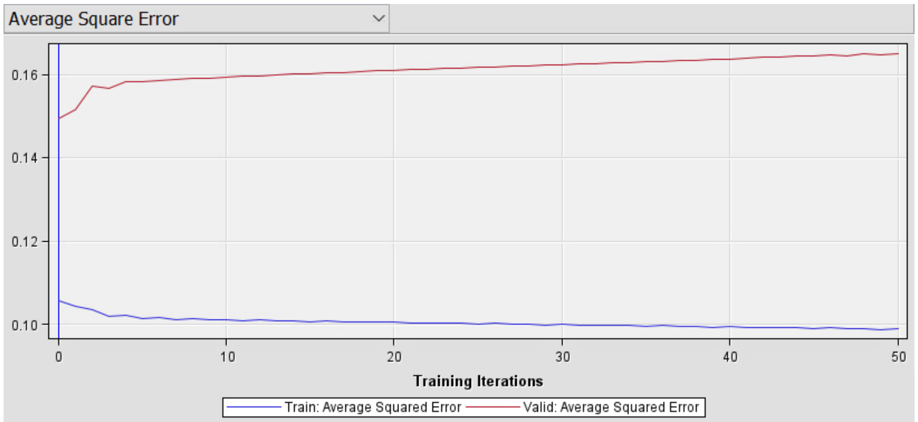
Training Technique	Default
Maximum Iterations	50
Maximum Time	30 Minutes
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1
Relative Function	0.0
Relative Function Times	1
Relative Gradient	1.0E-6
Relative Gradient Times	1
Propagation Options	
Accelerate	1.2
Decelerate	0.5
Learn	0.1
Maximum Learning	50.0
Minimum Learning	1.0E-5
Momentum	0.0
Maximum Momentum	1.75
Tilt	0.0
Preliminary Training	
Enable	Yes
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

4. Run the results:



Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of Freedom	459	.
DFE	Degrees of Freedom for Err...	416	.
DFM	Model Degrees of Freedom	43	.
NW	Number of Estimated Weig...	43	.
AIC	Akaike's Information Criterion	403.8618	.
SBC	Schwarz's Bayesian Criterion	581.4109	.
ASE	Average Squared Error	0.105768	0.149504
MAX	Maximum Absolute Error	0.979488	0.996727
DIV	Divisor for ASE	918	918
NOBS	Sum of Frequencies	459	459
RASE	Root Average Squared Error	0.32522	0.386658
SSE	Sum of Squared Errors	97.0951	137.2448
SUMW	Sum of Case Weights Time...	918	918
FPE	Final Prediction Error	0.127634	.
MSE	Mean Squared Error	0.116701	0.149504
RFPE	Root Final Prediction Error	0.357258	.
RMSE	Root Mean Squared Error	0.341615	0.386658
AVERR	Average Error Function	0.346255	0.478264
ERR	Error Function	317.8618	439.0464
MISC	Misclassification Rate	0.145969	0.20915
WRONG	Number of Wrong Classific...	67	96

The ASE for this model is 0.149504, the worst ASE out of all of our models thus far.

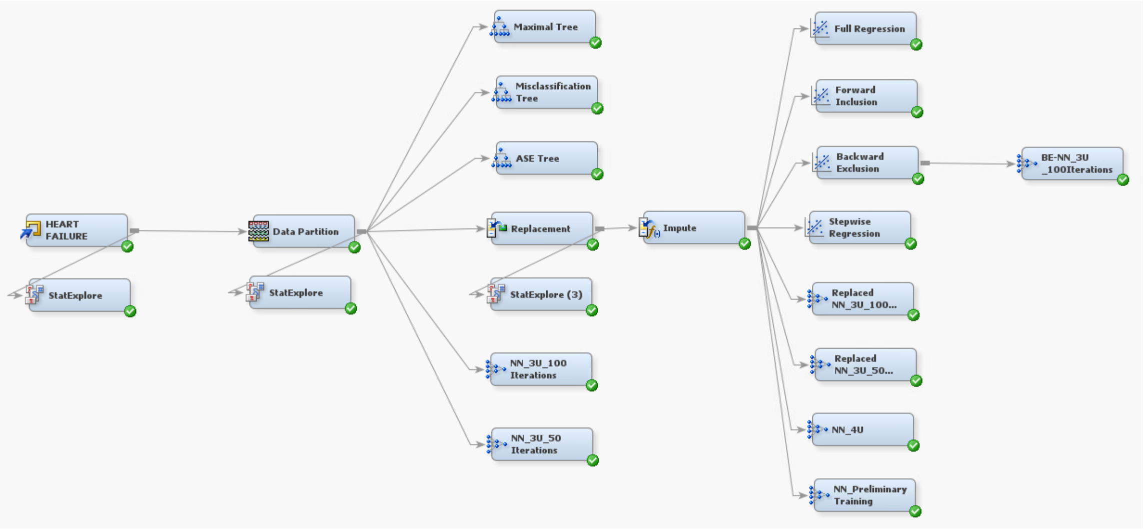


There is no convergence in this model.

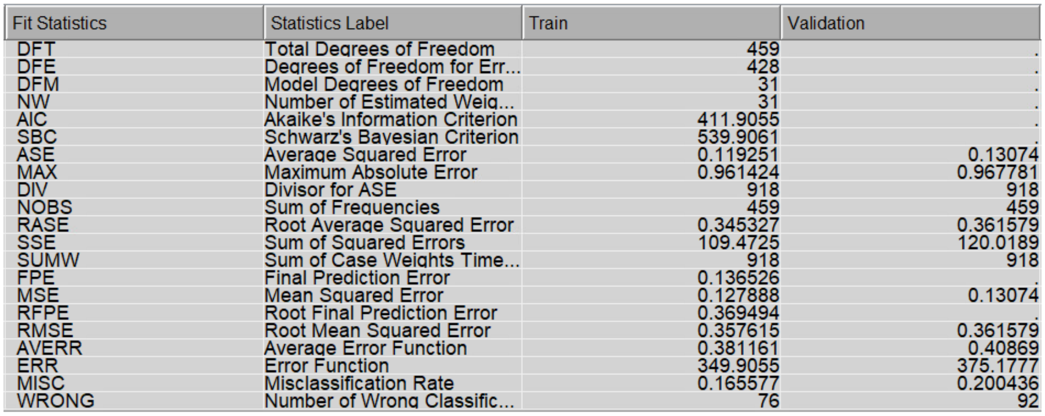
Neural Network with Backwards Regression (100 iterations):

Procedure:

1. Drag a **Neural Network** node into the diagram and connect it to the **Backward Exclusion** node.

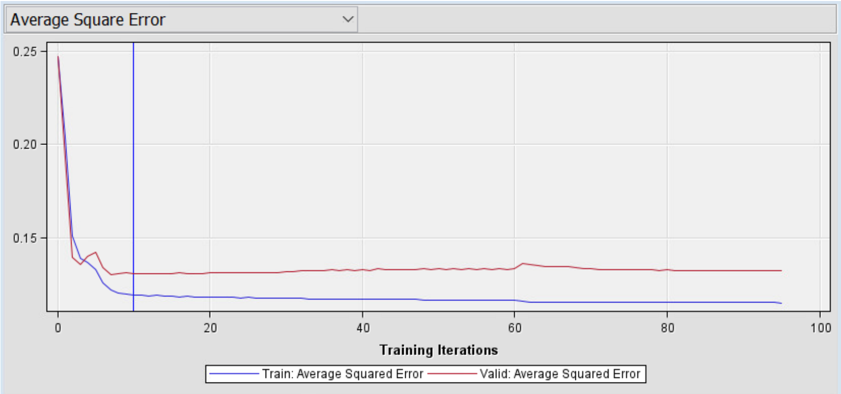


2. Select 3 Hidden Units and 100 iterations. Run the results:



Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of Freedom	459	.
DFE	Degrees of Freedom for Err...	428	.
DFM	Model Degrees of Freedom	31	.
NW	Number of Estimated Weig...	31	.
AIC	Akaike's Information Criterion	411.9055	.
SBC	Schwarz's Bayesian Criterion	539.9061	.
ASE	Average Squared Error	0.119251	0.13074
MAX	Maximum Absolute Error	0.961424	0.967781
DIV	Divisor for ASE	918	918
NOBS	Sum of Frequencies	459	459
RASE	Root Average Squared Error	0.345327	0.361579
SSE	Sum of Squared Errors	109.4725	120.0189
SUMW	Sum of Case Weights Time...	918	918
FPE	Final Prediction Error	0.136526	.
MSE	Mean Squared Error	0.127888	0.13074
RFPE	Root Final Prediction Error	0.369494	.
RMSE	Root Mean Squared Error	0.357615	0.361579
AVERR	Average Error Function	0.381161	0.40869
ERR	Error Function	349.9055	375.1777
MISC	Misclassification Rate	0.165577	0.200436
WRONG	Number of Wrong Classific...	76	92

The ASE for this model is 0.13074.

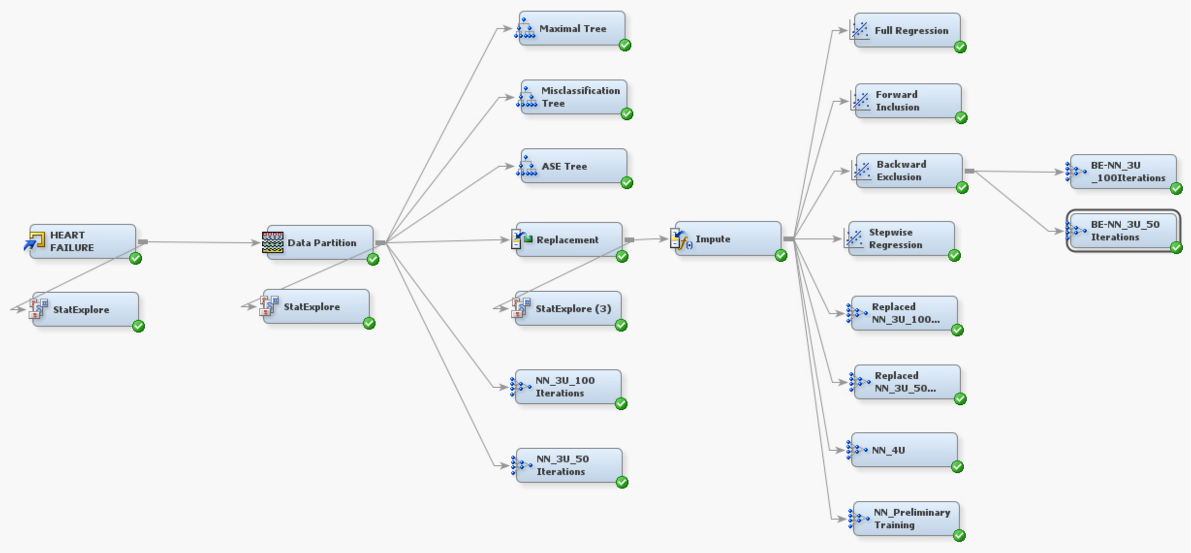


This model converges at 3 iterations.

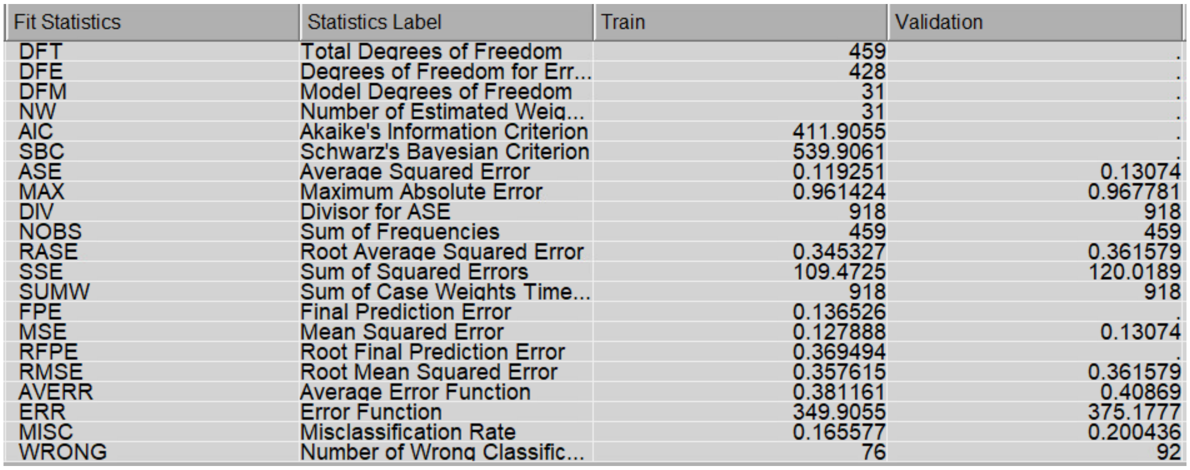
Neural Network with Backwards Regression (50 iterations):

Procedure:

1. Drag a **Neural Network** node into the diagram and connect it to the **Backward Exclusion** node:

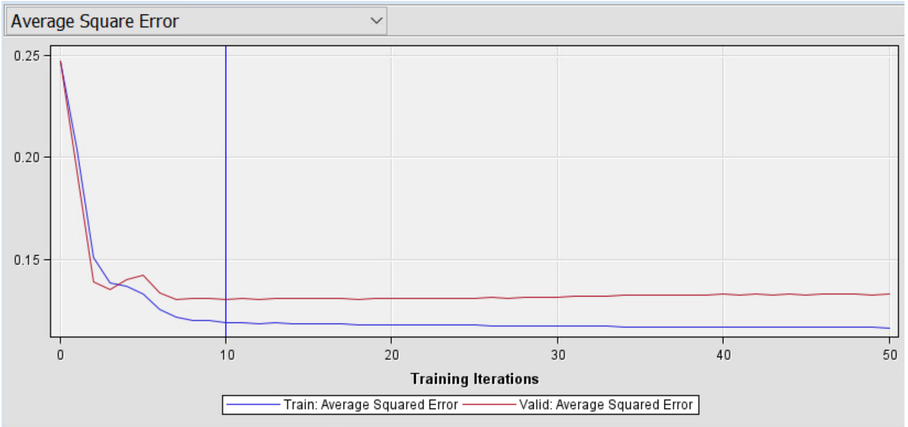


3. Select 3 Hidden Units and 50 iterations. Run the results:



Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of Freedom	459	.
DFE	Degrees of Freedom for Err...	428	.
DFM	Model Degrees of Freedom	31	.
NW	Number of Estimated Weig...	31	.
AIC	Akaike's Information Criterion	411.9055	.
SBC	Schwarz's Bayesian Criterion	539.9061	.
ASE	Average Squared Error	0.119251	0.13074
MAX	Maximum Absolute Error	0.961424	0.967781
DIV	Divisor for ASE	918	918
NOBS	Sum of Frequencies	459	459
RASE	Root Average Squared Error	0.345327	0.361579
SSE	Sum of Squared Errors	109.4725	120.0189
SUMW	Sum of Case Weights Time...	918	918
FPE	Final Prediction Error	0.136526	.
MSE	Mean Squared Error	0.127888	0.13074
RFPE	Root Final Prediction Error	0.369494	.
RMSE	Root Mean Squared Error	0.357615	0.361579
AVERR	Average Error Function	0.381161	0.40869
ERR	Error Function	349.9055	375.1777
MISC	Misclassification Rate	0.165577	0.200436
WRONG	Number of Wrong Classific...	76	92

This model has the same ASE as the 100 iteration model.



The model converges at 4 iterations.

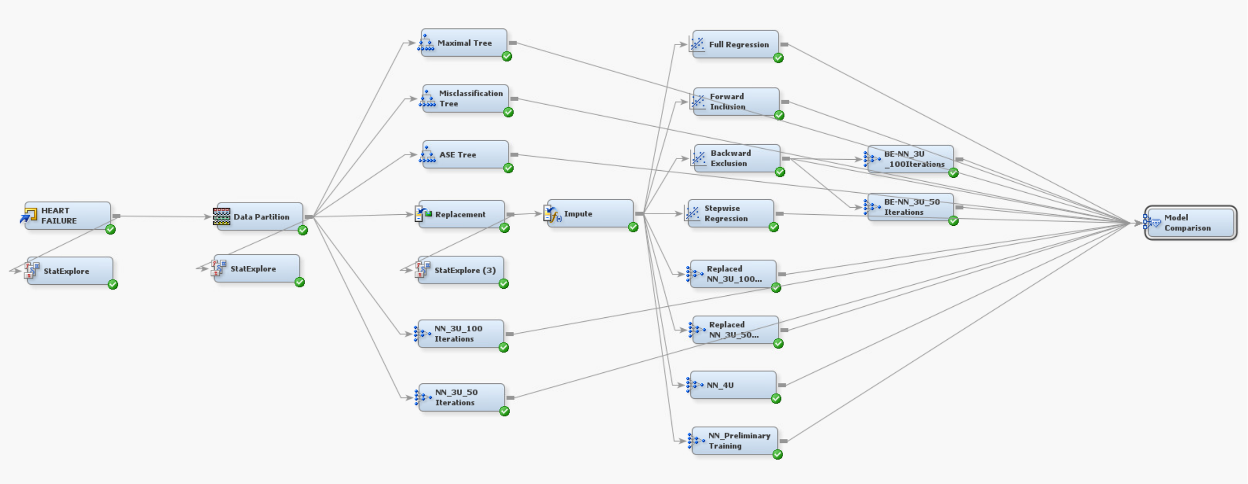
Conclusion:

Using ASE as our Model Selection method we can see that both of the Neural Network with 3 hidden units models are the best of our neural networks, as they have the lowest ASE.

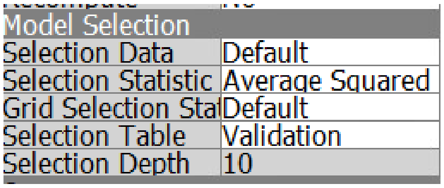
Model Comparison:

To determine which of our models has the best performance we ran a model comparison, using Validation ASE as the selection criterion.

1. Drag the **Model Comparison** node from the Assess tab into the diagram and connect all the models to the node.

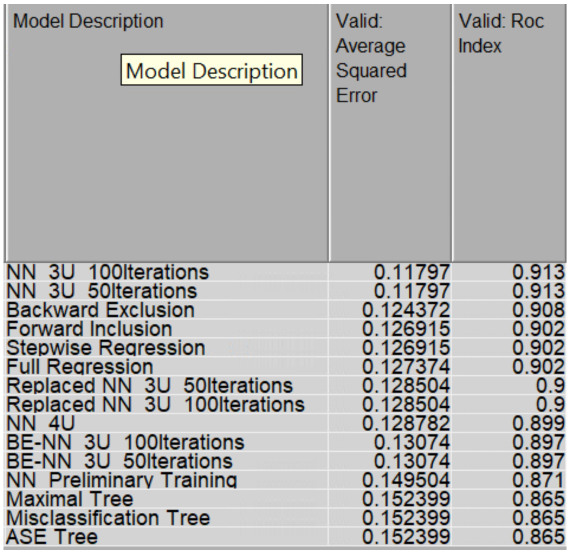


2. Set Selection Statistic to Average Squared Error and Selection Table to Validation.



Model Selection	
Selection Data	Default
Selection Statistic	Average Squared
Grid Selection Sta	Default
Selection Table	Validation
Selection Depth	10

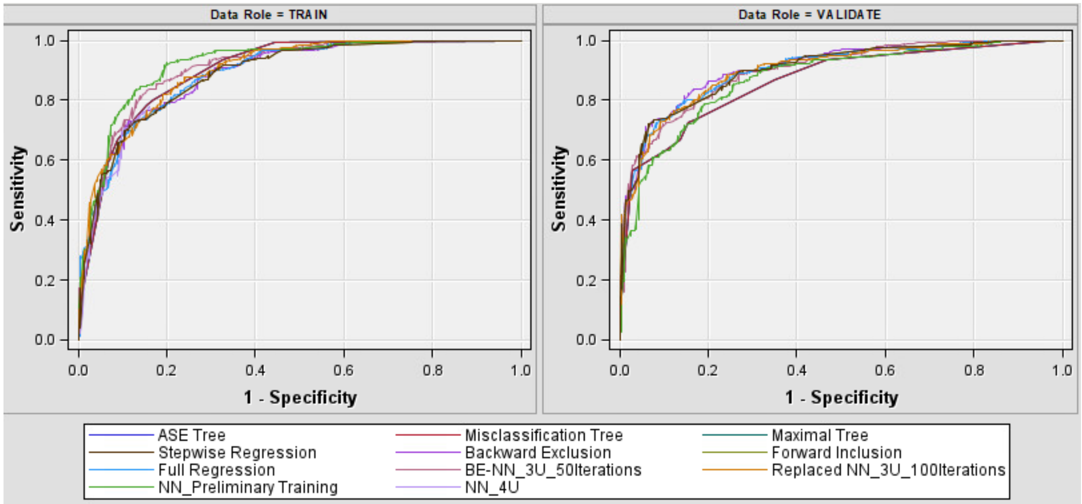
3. Run the node and open Fit Statistics. Drag the Roc Index next to the Selection Criterion.


Model Description

Model Description	Valid: Average Squared Error	Valid: Roc Index
NN 3U 100Iterations	0.11797	0.913
NN 3U 50Iterations	0.11797	0.913
Backward Exclusion	0.124372	0.908
Forward Inclusion	0.126915	0.902
Stepwise Regression	0.126915	0.902
Full Regression	0.127374	0.902
Replaced NN 3U 50Iterations	0.128504	0.9
Replaced NN 3U 100Iterations	0.128504	0.9
NN 4U	0.128782	0.899
BE-NN 3U 100Iterations	0.13074	0.897
BE-NN 3U 50Iterations	0.13074	0.897
NN Preliminary Training	0.149504	0.871
Maximal Tree	0.152399	0.865
Misclassification Tree	0.152399	0.865
ASE Tree	0.152399	0.865

Based on our Validation Criterion, the Replaced Neural Networks with 3 hidden units have the lowest Average Squared Error and the highest Roc Index, regardless of the number of iterations. However, these models have not been modified to exclude suspect values through replacement and imputation and are not reliable for this fact. Therefore, the best model is the Backward Exclusion Regression model. Further analysis will have to be conducted to determine why the models have better ASE and Roc Index scores when including erroneous values.

4. Expand the ROC Chart.



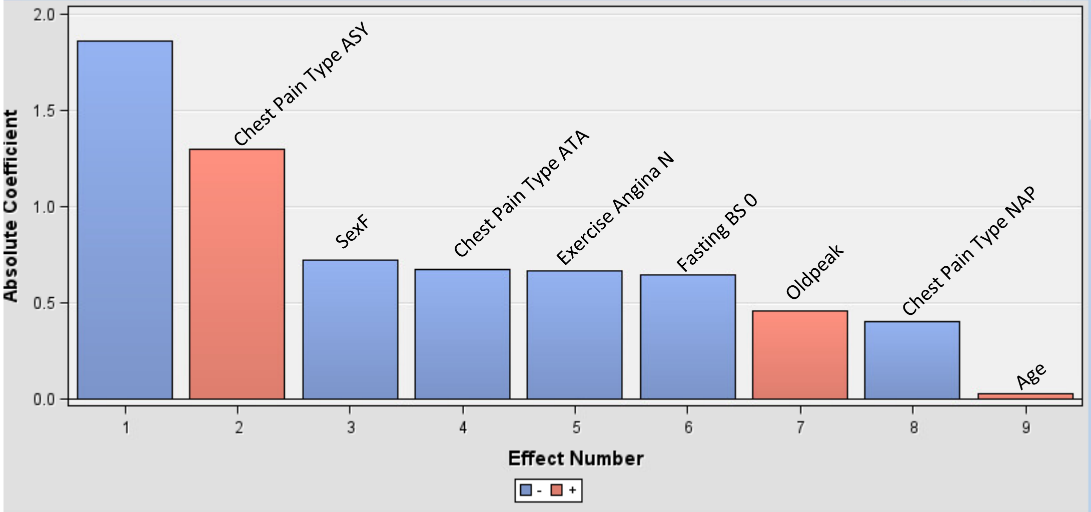
These models also have the highest Roc Curves.

Outcome:

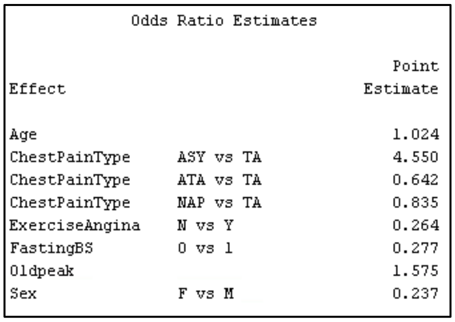
Based on our Model Comparison the best model for analyzing our data is the Backward Exclusion Regression Model. Of our reliable models, this model has the lowest ASE of 0.124372 and the highest Roc Index of 0.908.

Conclusion:

Now that we know our Backward Exclusion Regression is the best model, we can examine its results to determine which factors are most predictive of heart disease. To do this we will analyse the Odds Ratio Estimates and Effects Plot for the regression model.



The variables that are the strongest predictors of heart disease are as follows (descending order):



```
               Odds Ratio Estimates

                                          Point
Effect                                 Estimate

Age                                       1.024
ChestPainType     ASY vs TA               4.550
ChestPainType     ATA vs TA               0.642
ChestPainType     NAP vs TA               0.835
ExerciseAngina    N vs Y                  0.264
FastingBS         0 vs 1                  0.277
Oldpeak                                   1.575
Sex               F vs M                  0.237
```

1. Asymptomatic chest pain is 4.550 times more likely to be indicative of heart disease than Typical Angina chest pain.
2. For every unit higher of Oldpeak you are 1.57x more likely to experience heart disease.
3. For every year that you age your risk of contracting heart disease increases by 1.024x.
4. Typical Angina chest pain is 1.2x more indicative of heart disease than NAP chest pain.

5. Typical Angina chest pain is 1.6x more indicative of heart disease than Atypical Angina chest pain.
6. If your FastingBS levels are high, then you are 3.61x (1/0.277) more likely to suffer from heart disease.
7. Experiencing chest pain while exercising indicates that you are 3.8x (1/0.264) more likely to experience heart disease.
8. Males are 4.2x (1/0.237) more likely to suffer from heart disease than females.

Recommendations:

From our previous analysis, we have found that Asymptomatic chest pain is the most predictive variable for heart disease. Because the individual does not experience the regular symptoms that are indicators of heart disease (eg. chest pain, dizziness, nausea) the condition remains unidentified until it is too late, and the individual suffers from a "silent" heart attack or seizure that can result in their death. Our recommendations are therefore based on constant vigilance of one's health, regardless of symptoms.

1. Individuals should get regularly tested for heart disease, from the age of 28. Even if you are not currently experiencing symptoms this does not mean you are not at risk. Tests for heart disease should be administered during annual checkups.
2. Doctors should stress the importance of healthy diet and exercise to balance cholesterol and blood sugar levels. People should avoid excessive consumption of alcohol and the use of cigarettes, as these factors increase the likelihood of heart disease.
3. If you have suffered from a heart attack in the past your risk of contracting further heart disease is significantly higher. Individuals who have experienced a heart attack in the past should get routinely screened for any abnormalities that indicate heart disease.

Resources:

Fedesoriano. (2021, September 10). *Heart failure prediction dataset*. Kaggle. Retrieved December 17, 2021, from https://www.kaggle.com/fedesoriano/heart-failure-prediction