<u>OMIS 482 PROJECT 2</u>

Kevin Rius

Jeff Willer

<u>NOTE: Again, please include screenshots of any major steps and an explanation behind each thing you did. You will need to present this so you will also need to create a PowerPoint presentation.**</u>**

1. Find a data set that has at least 1000 observations and at least 10 input variables. **Please get the data set approved by me prior to moving on with the project. This is also a first come, first serve type of situation. If someone else finds the same data set first and asks me about it, it cannot be used by another group.** This does not need to be a SAS data file, as you can convert many data types to SAS files. If you need help with this step, please ask. Some data file types are easier to convert using the SAS Studio software and I can always do this conversion for you if necessary. You may also find your data as split, multiple data files. These can be combined and used if it is something you are really interested in using.

   Below is our dataset file we used for our analysis!

   cwurDataProject2.xls

2. Please describe the data to me in your project. Please tell me if the target is binary or continuous and describe all of the input variables and their measurement scales. Why are you using predictive modeling techniques on this data?

Our Variables:

VARIABLES

- World_Rank = world rank for each university *(Interval)* ***REJECTED***
- Institution = name of university *(Nominal)*
- Country = country of each university *(Nominal)*
- National_Rank = rank of university within its country *(Interval)*
- Quality_of_Education = rank for quality of education *(Interval)*
- Alumni_Employment = rank for alumni employment *(Interval)*
- Quality_of_Faculty = rank for quality of faculty *(Interval)*
- Publications = rank for publications *(Interval)*
- Influence = rank for influence on students & their educational growth *(Interval)*
- Citations = number of students at the university *(Interval)*
- Broad_Impact = rank for broad impact (only for 2014-2015) *(Interval)* ***DROPPED***
- Patents = rank for patents from University *(Interval)*
- Score = total score, used for determining world rank; highest score = greatest rank *(Interval)* ***TARGET***
- Year = year of ranking (2012 to 2015) *(Interval)*

Variables - FIMPORT

(none) | not | Equal to | ...

Columns: Label | Mining | Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
alumni_employn	Input	Interval	No		No	.	.
citations	Input	Interval	No		No	.	.
country	Input	Nominal	No		No	.	.
influence	Input	Interval	No		No	.	.
institution	Input	Nominal	No		No	.	.
national_rank	Input	Interval	No		No	.	.
patents	Input	Interval	No		No	.	.
publications	Input	Interval	No		No	.	.
quality_of_educ	Input	Interval	No		No	.	.
quality_of_facul	Input	Interval	No		No	.	.
score	Target	Interval	No		No	.	.
world_rank	Rejected	Interval	No		No	.	.
year	Input	Interval	No		No	.	.

- **Our target variable is Score (0.00-100.00) which is an interval variable. World_Rank was rejected because World_rank and score have a very similar relationship. The university with the highest score will also have the highest rank and vice-versa, so we decided to take the variable out because it would affect variable worth and other statistical measures to determine the factors for the best university. We are using predictive modeling techniques on this data to determine which variables influence having the "best" university, meaning which variables have the biggest contribution and impact on the score which gives each university its rank. We would like to find out according to the data collected which factors contribute to making a great university and educational institution!**

3. Create a new project in SAS.
4. Create a new library in SAS that links to whatever folder your data is saved in.
5. Create a new diagram in SAS.
6. Input the data set into SAS. Make sure the target is set to be a target variable and all measurement scales are correct. Put the data set on your diagram.
7. Do data partition. You can choose the amount you allocate to each one but please tell me what you chose. This can be influenced by your sample size so if you have any concerns with what you set up, please ask me before proceeding.

- **We decided to use the default values for data partition: Training = 40.0, Validation = 30.0, and Test = 30.0**

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	40.0
Validation	30.0
Test	30.0
Report	
Interval Targets	Yes
Class Targets	No

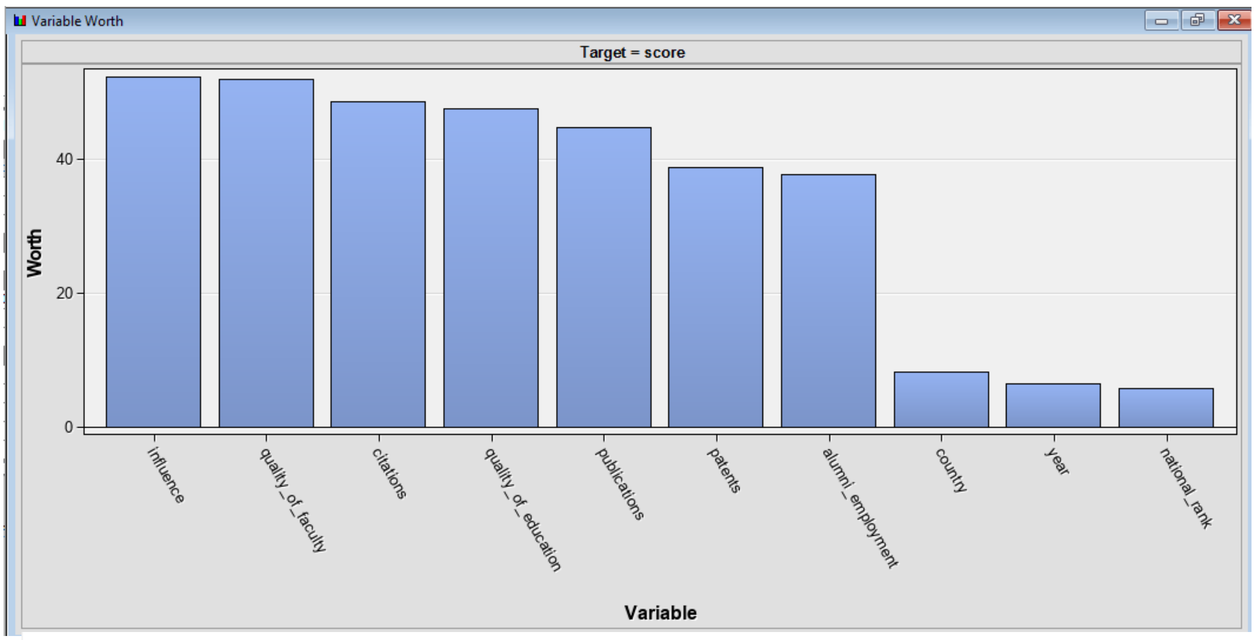
8. Go through all the **sample nodes** we went over in chapter 2 (Input Data, Data Partition, Filter, File Import, Time Series, Merge, and Append). Does your data require you to use any of these nodes (other than input data or data partition)? If so, please add it onto the diagram and explain why the node was added and what was done in terms of settings. You can also explain why you chose NOT to use any of these nodes.

    - **The only sample nodes we used in our analysis was the File Import node and Data Partition node. We need to use the File Import node to input our dataset into SAS EM for analysis and we also were required to use the Data Partition node for further analysis. We did not use any other sample nodes we went over in chapter 2 because we felt our data set did not need these nodes to be applied. The Time Series node was not very necessary for our data as we only had year dates which were not very impactful on the data to begin with, the Merge and Append nodes were also not needed for our dataset we used in our analysis.**

9. Go through all the **initial data exploration** nodes we went over in chapter 2 (Stat Explore, MultiPlot, Graph Explore, Variable Clustering, and Cluster). You should run Stat Explore and MultiPlot to look at the distributions of your variables and comment on what you see. Would it make sense to use any of the other nodes from this section for your data? If so, please add it onto the diagram and explain why the node was added and what was done in terms of settings.

- **We decided to use both the StatExplore and Multiplot nodes as the only initial data exploration nodes. Below you can view the results of the Stat Explore Node and Multiplot Node. You can see in the Variable Worth chart the variables with the highest worth (variables with the highest worth are the most important variables in relation to the target variable – score) Variable worth is calculated from the p-values of the chi-square statistics! When using the StatExplore node for a continuous target the interval variables property must be set to No, and the Correlations, Pearson Correlations and Spearman Correlations properties are all set to Yes.**

Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	No
Number of Bins	5
Correlation Statistics	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	Yes
Status	



Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
alumni_employment	INPUT	358.9932	187.5072	880	0	1	456	567	-0.51704	-1.21521
citations	INPUT	409.6045	261.0195	880	0	1	428	812	0.072282	-1.23329
influence	INPUT	459.2091	304.6086	880	0	1	441	991	0.116506	-1.2872
national_rank	INPUT	40.69659	52.1435	880	0	1	21	229	1.963456	3.171855
patents	INPUT	435.3977	275.9086	880	0	1	426	871	0.005226	-1.36624
publications	INPUT	460.0739	304.0745	880	0	1	442	999	0.106756	-1.27246
quality_of_education	INPUT	273.775	122.273	880	0	1	355	367	-0.97532	-0.6815
quality_of_faculty	INPUT	179.092	64.39684	880	0	1	210	218	-1.56055	0.895544
year	INPUT	2014.35	0.755864	880	0	2012	2014	2015	-1.28069	0.583087
score	TARGET	47.90466	7.950473	880	0	43.36	45.11	100	4.002199	18.11259

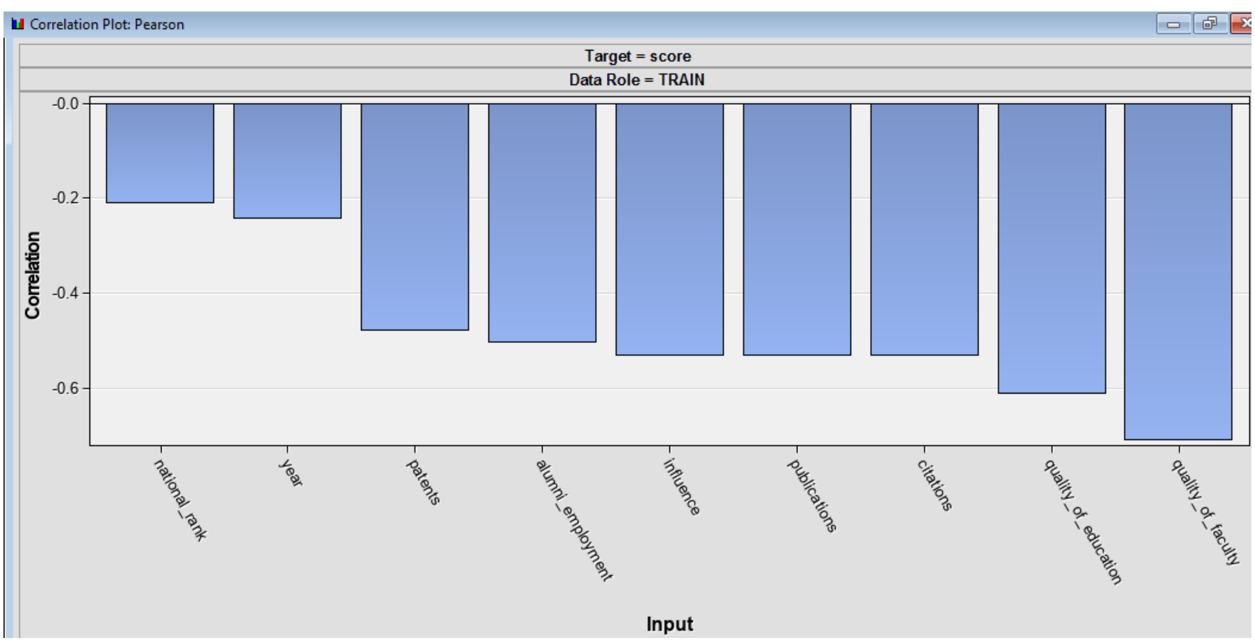
Correlation Statistics
(maximum 500 observations printed)

Data Role=TRAIN Type=PEARSON Target=score

Input	Correlation
national_rank	-0.20832
year	-0.24161
patents	-0.47717
alumni_employment	-0.50227
influence	-0.52971
publications	-0.52974
citations	-0.53141
quality_of_education	-0.61116
quality_of_faculty	-0.70702

Data Role=TRAIN Type=SPEARMAN Target=score

Input	Correlation
year	-0.20763
national_rank	-0.31208
quality_of_faculty	-0.57137
alumni_employment	-0.57340
quality_of_education	-0.60081
patents	-0.64924
citations	-0.80715
influence	-0.84104
publications	-0.85368



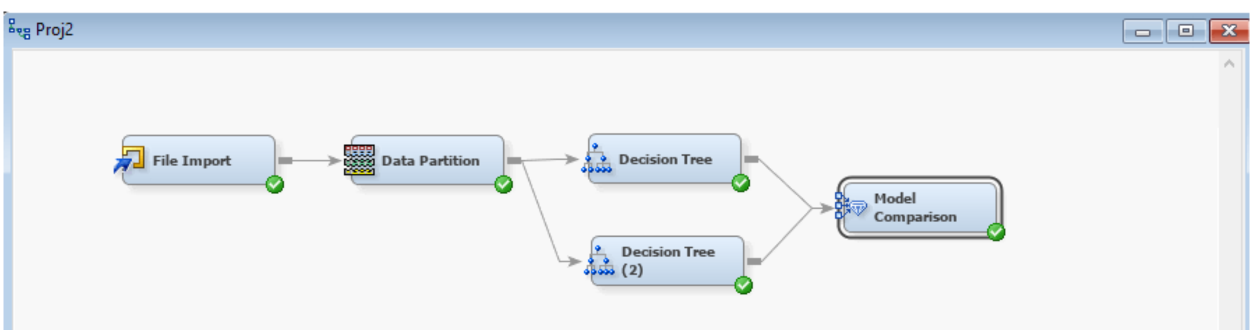
10. Go through all the **tools for data modification** nodes we went over in chapter 2 (Drop, Replacement, Impute, Interactive Binning, and Principal Components). Would it make sense to use any of these nodes for your data? If so, please add it onto the diagram and explain why the node was added and what was done.

    - **We did not need to use any tools for data modification nodes within our analysis.**

11. Are you choosing to do variable selection or transformation of variables? What is your reasoning behind your decision? If you do choose to do either of these, add the nodes to the diagram and set the appropriate settings.

    - **Prior to adding our data set into SASEM we removed one variable (broad_impact as it was a variable only collected for 2 of the 4 years within the data set and that is why we removed it. Additionally, we rejected the World_rank variables as prior explained!) We did not want any other variables to end up being rejected with a low R-squared value as we wanted to see each and every variable provided in the dataset in our analysis.**

12. Add some decision tree models to the diagram and vary the settings in each one to produce different trees. The number of trees you can create can depend on the type of target variable you chose. Explain why you built the trees you built.



    - **We decided to use two decision tree nodes and the reason we only did two is because with our target variable being interval we were able to selection two interval target criterion. The first decision tree used ProbF while decision tree 2 used Variance as the criterion.**

Property	Value
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No

Property	Value
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	Variance
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	

- **We connected a model comparison node to each decision tree node to determine which model would be best. After running the model comparison node, it was determined that decision tree 2 (variance criterion) was the better of the two decision tree models. Decision tree 2 had the lower Average Square Error and was selected. The results can be viewed below:**

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom
Y	Tree2	Tree2	Decision Tr...	score	score	4.376768	880	12.10083	2183.747	2.481531	1.575287	880	88
	Tree	Tree	Decision Tr...	score	score	4.724542	880	12.10083	2968.662	3.373479	1.836703	880	88

Test: Sum of Frequencies	Test: Sum of Weights Times Freqs	Test: Maximum Absolute Error	Test: Sum of Squared Errors	Test: Average Squared Error	Test: Root Average Squared Error	Test: Divisor for TASE
660	660	19.96917	3451.885	5.230129	2.286948	660
660	660	19.96917	3505.902	5.311973	2.304772	660

```
Data Role=Test

           Statistics                        Tree2       Tree

Test: Average Squared Error                   5.23       5.31
Test: Divisor for TASE                      660.00     660.00
Test: Maximum Absolute Error                 19.97      19.97
Test: Sum of Frequencies                    660.00     660.00
Test: Root Average Squared Error              2.29       2.30
Test: Sum of Squared Errors                3451.89    3505.90
Test: Sum of Weights Times Freqs            660.00     660.00
```

13. Add different types of neural network models to the diagram and vary the settings to produce different models. Explain why you chose to add each type of neural network model to the diagram.

- **We only used one neural network node within our diagram due to having an interval target variable. The settings used for our neural network node can be viewed below. We used Average Error as our Model Selection Criterion. Additionally, we used Multilayer Perception as our Architecture, and for our Target Layer Activation Function we used the identity function since our target variable is interval!**

Property	Value
General	
Node ID	Neural
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection Criterion	Average Error
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	

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	Identity
Target Layer Error Function	Default

- **Here are the Fit Statistics results from the Neural Network Node:**

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
score	score	_DFT_	Total Degre...	880	.	.
score	score	_DFE_	Degrees of ...	-1263	.	.
score	score	_DFM_	Model Degr...	2143	.	.
score	score	_NW_	Number of ...	2143	.	.
score	score	_AIC_	Akaike's Inf...	.	.	.
score	score	_SBC_	Schwarz's ...	.	.	.
score	score	_ASE_	Average Sq...	4.807647	22.19633	22.40179
score	score	_MAX_	Maximum A...	19.44566	49.50534	48.23534
score	score	_DIV_	Divisor for ...	880	660	660
score	score	_NOBS_	Sum of Fre...	880	660	660
score	score	_RASE_	Root Avera...	2.192635	4.711298	4.733053
score	score	_SSE_	Sum of Squ...	4230.73	14649.58	14785.18
score	score	_SUMW_	Sum of Cas...	880	660	660
score	score	_FPE_	Final Predic...	.	.	.
score	score	_MSE_	Mean Squa...	.	22.19633	22.40179
score	score	_RFPE_	Root Final ...	.	.	.
score	score	_RMSE_	Root Mean ...	.	4.711298	4.733053
score	score	_AVERR_	Average Err...	4.807647	22.19633	22.40179
score	score	_ERR_	Error Functi...	4230.73	14649.58	14785.18
score	score	_MISC_	Misclassific...	.	.	.
score	score	_WRONG_	Number of ...	.	.	.

14. Add different types of regression models to the diagram. Vary the selection model property to build the different models. Make sure the regression type property is set correctly for your type of target variable.

- **We used 4 different types of regression models in our diagram varying the selection model property to build different models. The regression type property was set to Linear Regression due to having an interval target! The 4 different types of regression models used were as followed:**

Regression Node 1: Selection Model = None, Selection Criterion = Default

.. Property	Value
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour

Regression Node 2: Selection Model = Stepwise, Selection Criterion = Validation Error

.. Property	Value
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour

Regression Node 3: Selection Model = Forward, Selection Criterion = Validation Error

Property	Value
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour

Regression Node 4: Selection Model = Backward, Selection Criterion = Validation Error

Property	Value
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour

- **All 4 regression nodes were connected to a Model Comparison node to view which regression model was the best out of all of them. The first regression node was selected as the best model (Selection Model = None, Selection Criterion = Default) this model had the lowest Test Average Square Error and was selected. The results of all the four regression nodes can be viewed below:**

<u>Fit Statistics:</u>

Fit Statistics

alid: Sum f Case Weights imes Freq	Test: Average Squared Error	Test: Lower 95% Conf. Limit for TASE	Test: Upper 95% Conf. Limit for TASE	Test: Average Error Function	Test: Divisor for TASE	Test: Error Function	Test: Maximum Absolute Error	Test: Mean Square Error	Test: Sum of Frequencies	Test: Root Average Squared Error	Test: Root Mean Square Error	Test: Sum of Square Errors	Test: Sum of Case Weights Times Freq
660	21.36013	11.79958	33.71077	21.36013	660	14097.68	48.23534	21.36013	660	4.621702	4.621702	14097.68	660
660	21.36013	11.79958	33.71077	21.36013	660	14097.68	48.23534	21.36013	660	4.621702	4.621702	14097.68	660
660	21.38495	11.80449	33.76483	21.38495	660	14114.07	48.23534	21.38495	660	4.624387	4.624387	14114.07	660
660	21.56467	11.85913	34.12406	21.56467	660	14232.68	48.23534	21.56467	660	4.643778	4.643778	14232.68	660

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

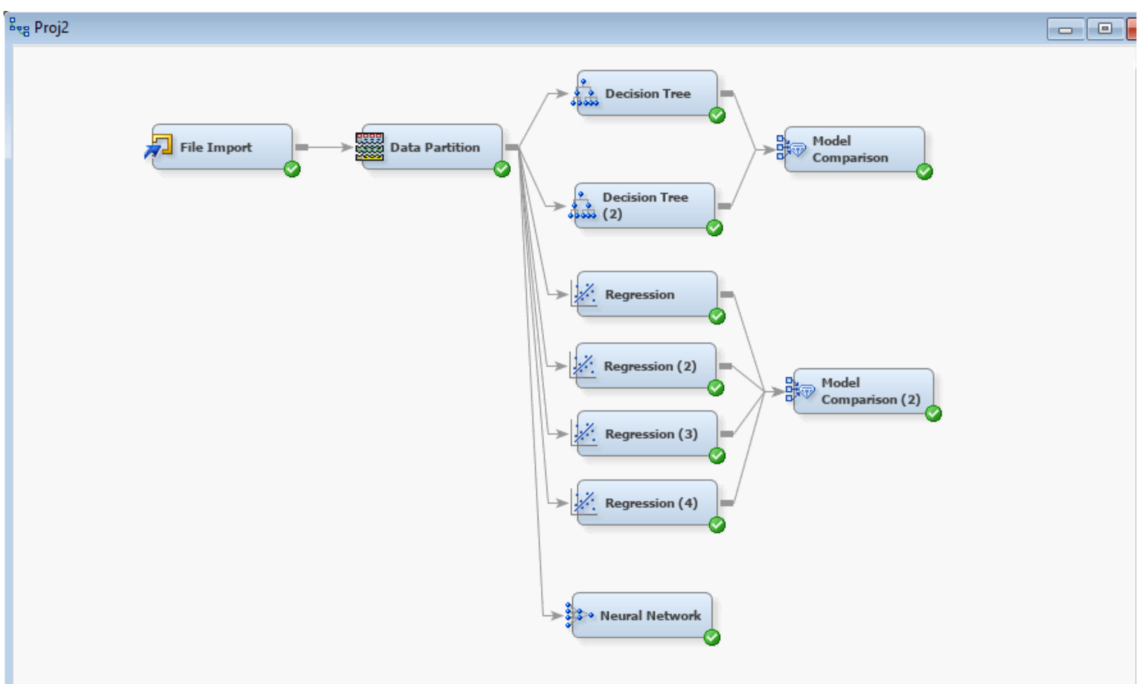
Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Reg	Regression	21.4252	1.23616
	Reg4	Regression (4)	21.4252	1.23616
	Reg3	Regression (3)	21.5051	1.23922
	Reg2	Regression (2)	22.1258	1.55025

Test Statistics:

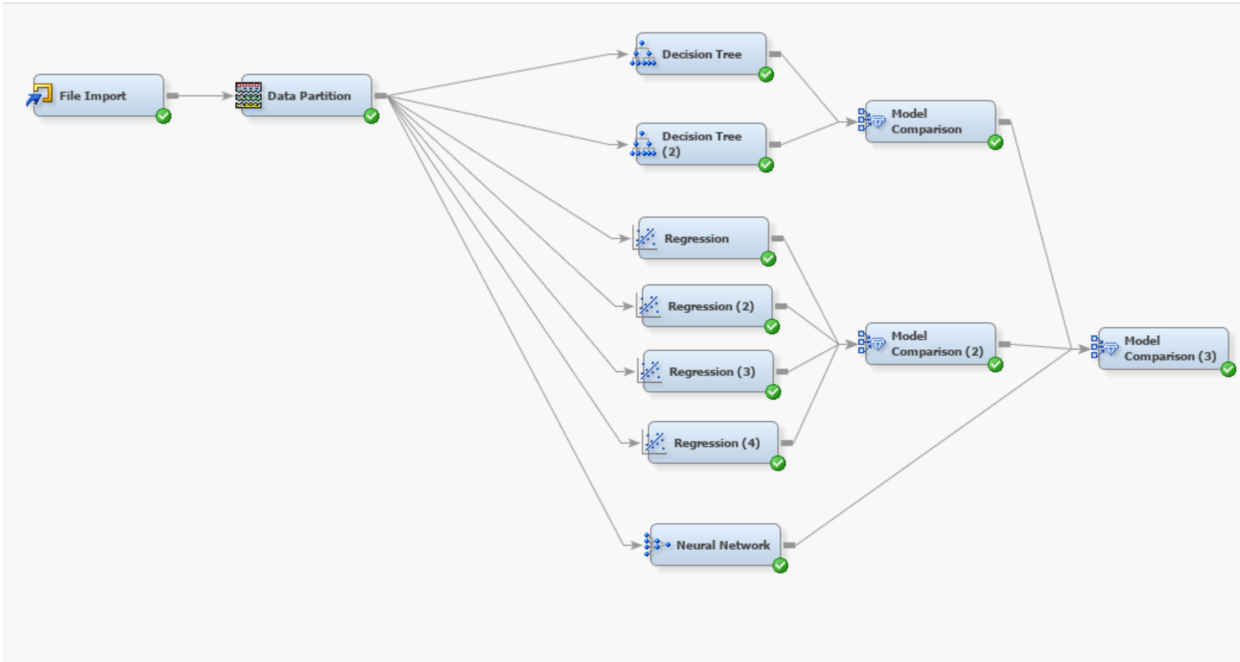
Data Role=Test

Statistics	Reg	Reg4	Reg3	Reg2
Test: Lower 95% Conf. Limit for TASE	11.80	11.80	11.80	11.86
Test: Upper 95% Conf. Limit for TASE	33.71	33.71	33.76	34.12
Test: Average Squared Error	21.36	21.36	21.38	21.56
Test: Average Error Function	21.36	21.36	21.38	21.56
Test: Divisor for TASE	660.00	660.00	660.00	660.00
Test: Error Function	14097.68	14097.68	14114.07	14232.68
Test: Maximum Absolute Error	48.24	48.24	48.24	48.24
Test: Mean Square Error	21.36	21.36	21.38	21.56
Test: Sum of Frequencies	660.00	660.00	660.00	660.00
Test: Root Average Squared Error	4.62	4.62	4.62	4.64
Test: Root Mean Square Error	4.62	4.62	4.62	4.64
Test: Sum of Square Errors	14097.68	14097.68	14114.07	14232.68
Test: Sum of Case Weights Times Freq	660.00	660.00	660.00	660.00

COMPLETED DIAGRAM BEFORE COMPARING ALL MODELS:



15. Add a model comparison node and use different criteria to explain which model ends up being the best one.



- **After running the model comparison node combining the best Decision Tree (2), the best Regression model, and the lone Neural Network model, the Decision Tree (2) was selected due to having the lowest Average Square Error (Test data) of 5.23 compared to the Regression model having 21.36 (Test data), and the Neural Network model having an Average Square error of 22.40179 (Test data). The results of the final model comparison node can be viewed below:**

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE
Y	MdlComp	Tree2	Decision Tr...	score	score	4.376768	.	2.481531	.	.	.	880	88
	MdlComp2	Reg	Regression	score	score	21.42515	1506.567	1.236158	1.236158	220	660	880	88
	Neural	Neural	Neural Net...	score	score	22.19633	.	4.807647	4.807647	-1263	2143	880	88

Test: Average Squared Error	Test: Lower 95% Conf. Limit for TASE	Test: Upper 95% Conf. Limit for TASE	Test: Average Error Function	Test: Divisor for TASE	Test: Error Function	Test: Maximum Absolute Error	Test: Mean Square Error	Test: Sum of Frequencies	Test: Root Average Squared Error	Test: Root Mean Square Error	Test: Sum of Square Errors	Test: Sum of Case Weights Times Freq	Train: Misclassification Rate
5.230129	.	.	.	660	.	19.96917	.	660	2.286948	.	3451.885	660	
21.36013	11.79958	33.71077	21.36013	660	14097.68	48.23534	21.36013	660	4.621702	4.621702	14097.68	660	
22.40179	.	.	22.40179	660	14785.18	48.23534	22.40179	660	4.733053	4.733053	14785.18	660	

Data Role=Test

Statistics	Tree2	Reg	Neural
	.	.	.
	.	.	.
Test: Lower 95% Conf. Limit for TASE	.	11.80	.
Test: Upper 95% Conf. Limit for TASE	.	33.71	.
Test: Average Squared Error	5.23	21.36	22.40
Test: Average Error Function	.	21.36	22.40
Test: Divisor for TASE	660.00	660.00	660.00
Test: Error Function	.	14097.68	14785.18
	.	.	.
	.	.	.
	.	.	.
	.	.	.
Test: Maximum Absolute Error	19.97	48.24	48.24
Test: Misclassification Rate	.	.	.
Test: Lower 95% Conf. Limit for TMISC	.	.	.
Test: Upper 95% Conf. Limit for TMISC	.	.	.
Test: Mean Square Error	.	21.36	22.40
Test: Sum of Frequencies	660.00	660.00	660.00
Test: Root Average Squared Error	2.29	4.62	4.73
Test: Root Mean Square Error	.	4.62	4.73
Test: Sum of Square Errors	3451.89	14097.68	14785.18
Test: Sum of Case Weights Times Freq	660.00	660.00	660.00
Test: Number of Wrong Classifications	.	.	.

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