Data Mining Team Project

PTSD Team

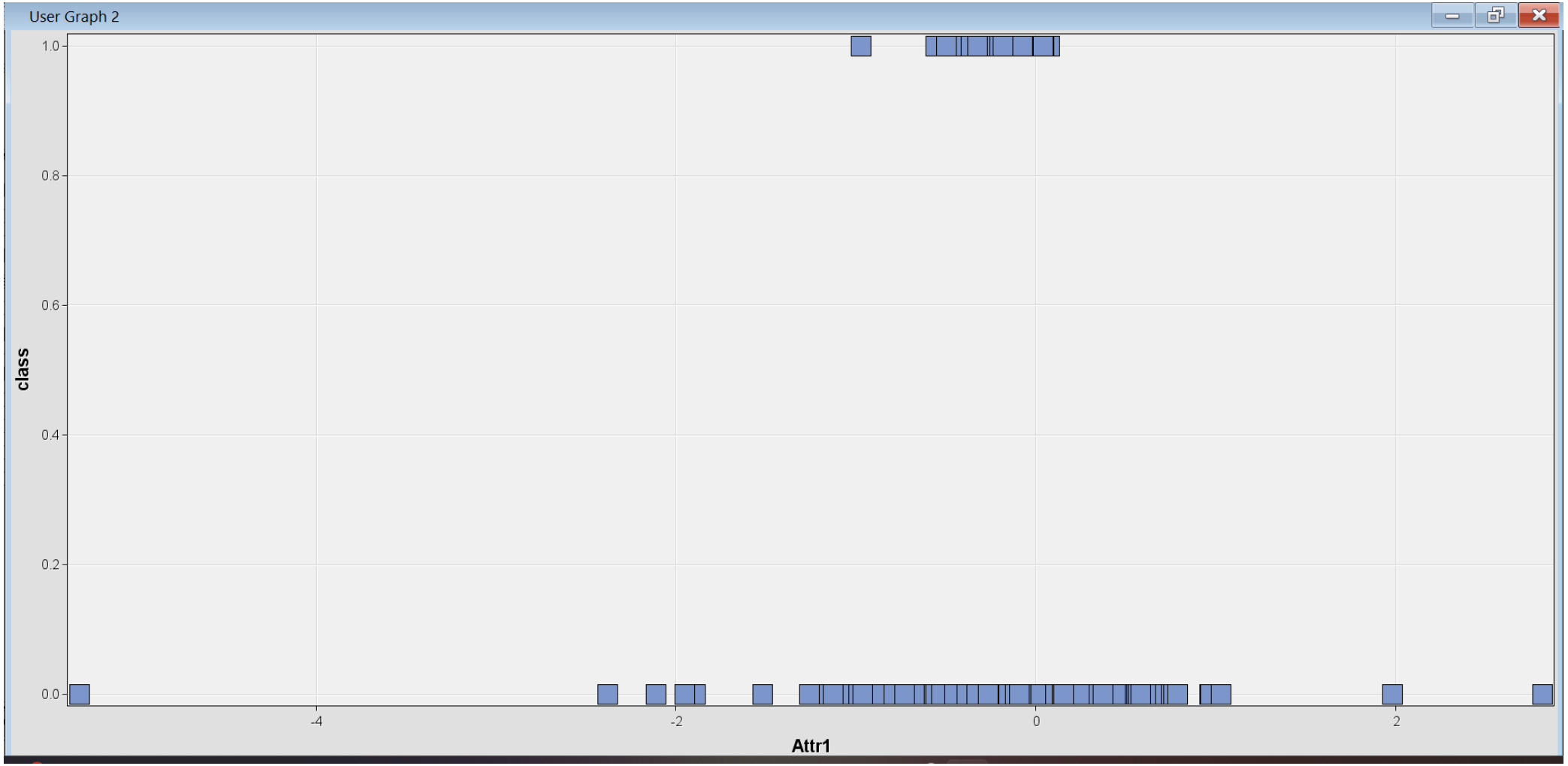
Pooja Udayanjali Kannuri

Shubhankar D Bhajekar

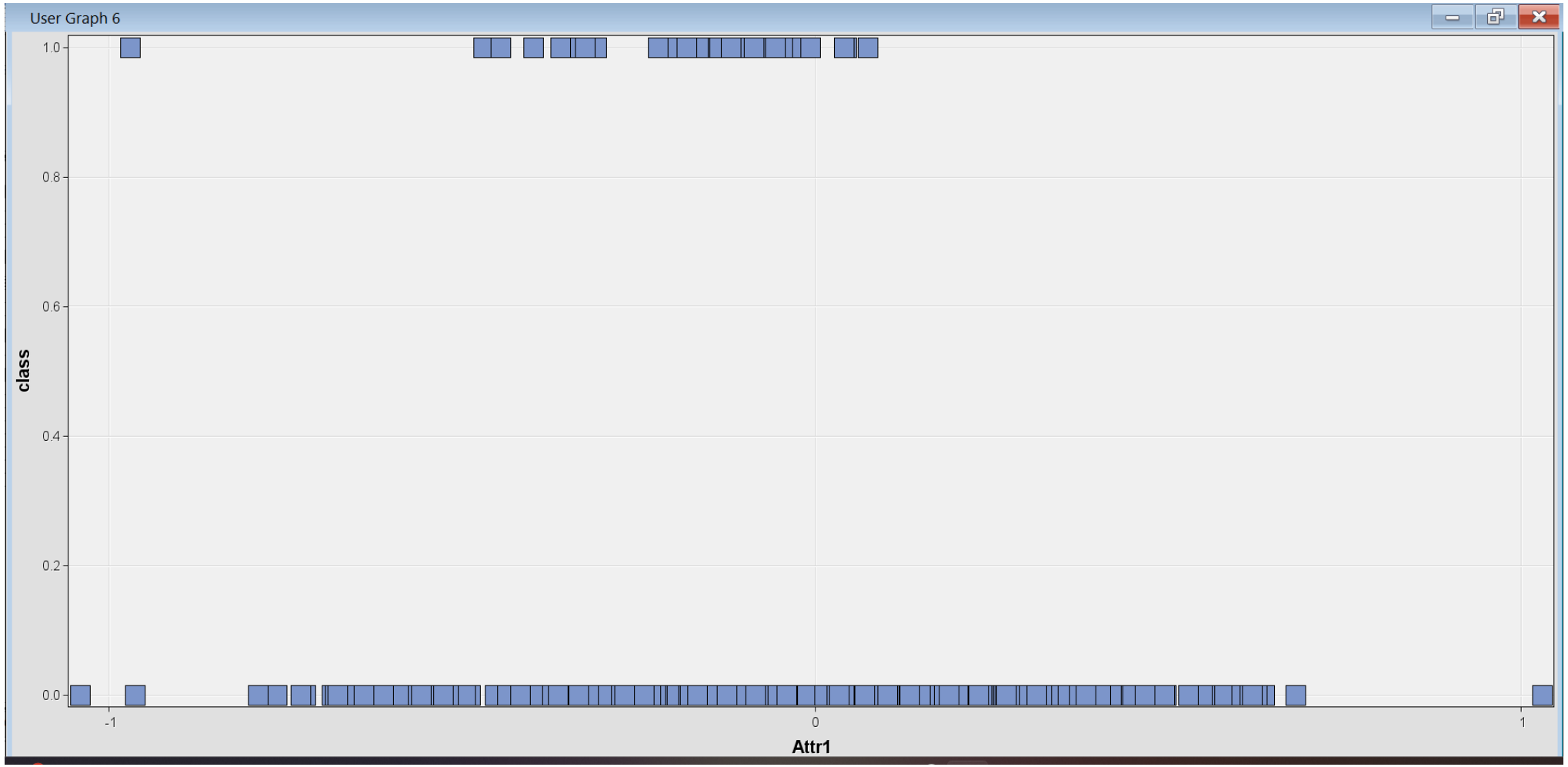
Theoni Dounia

PURDUE UNIVERSITY®

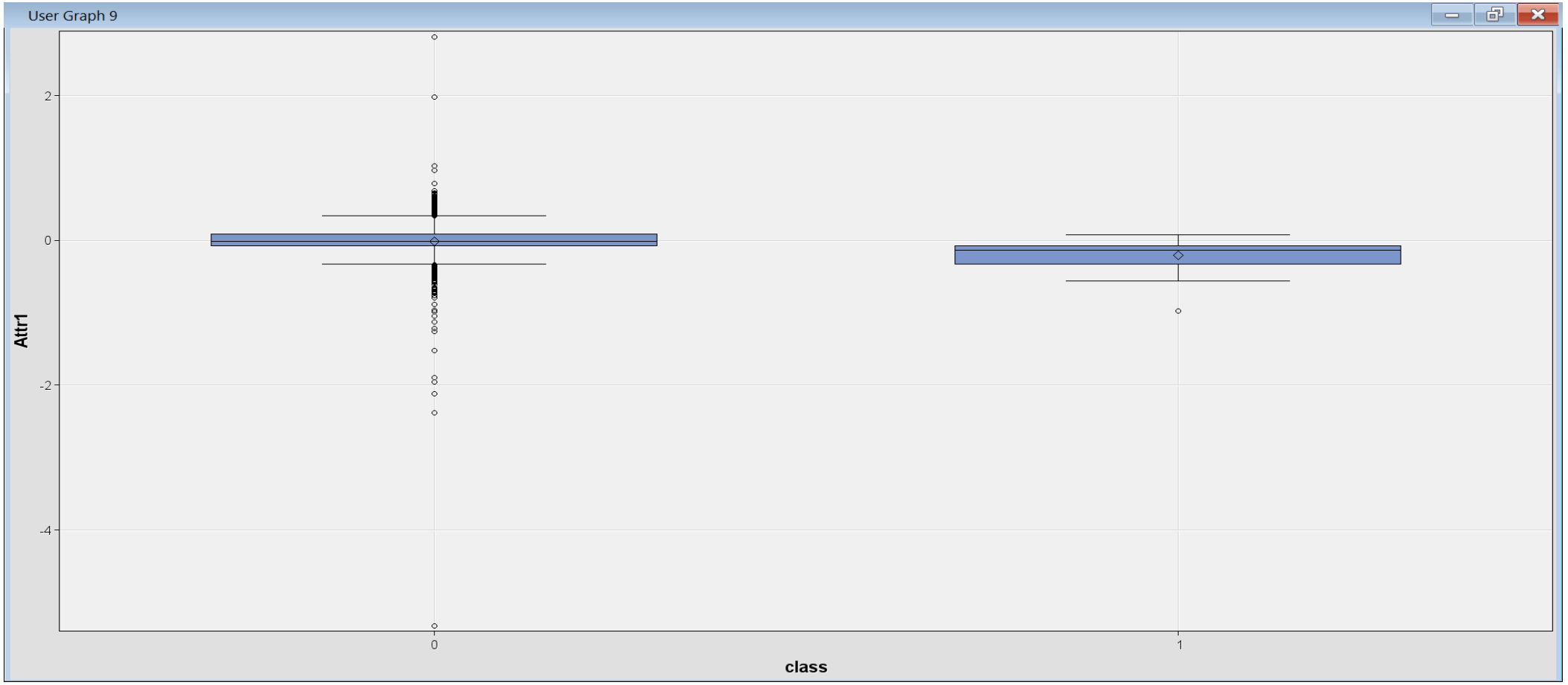
Attribute 1 – Before Filter Node



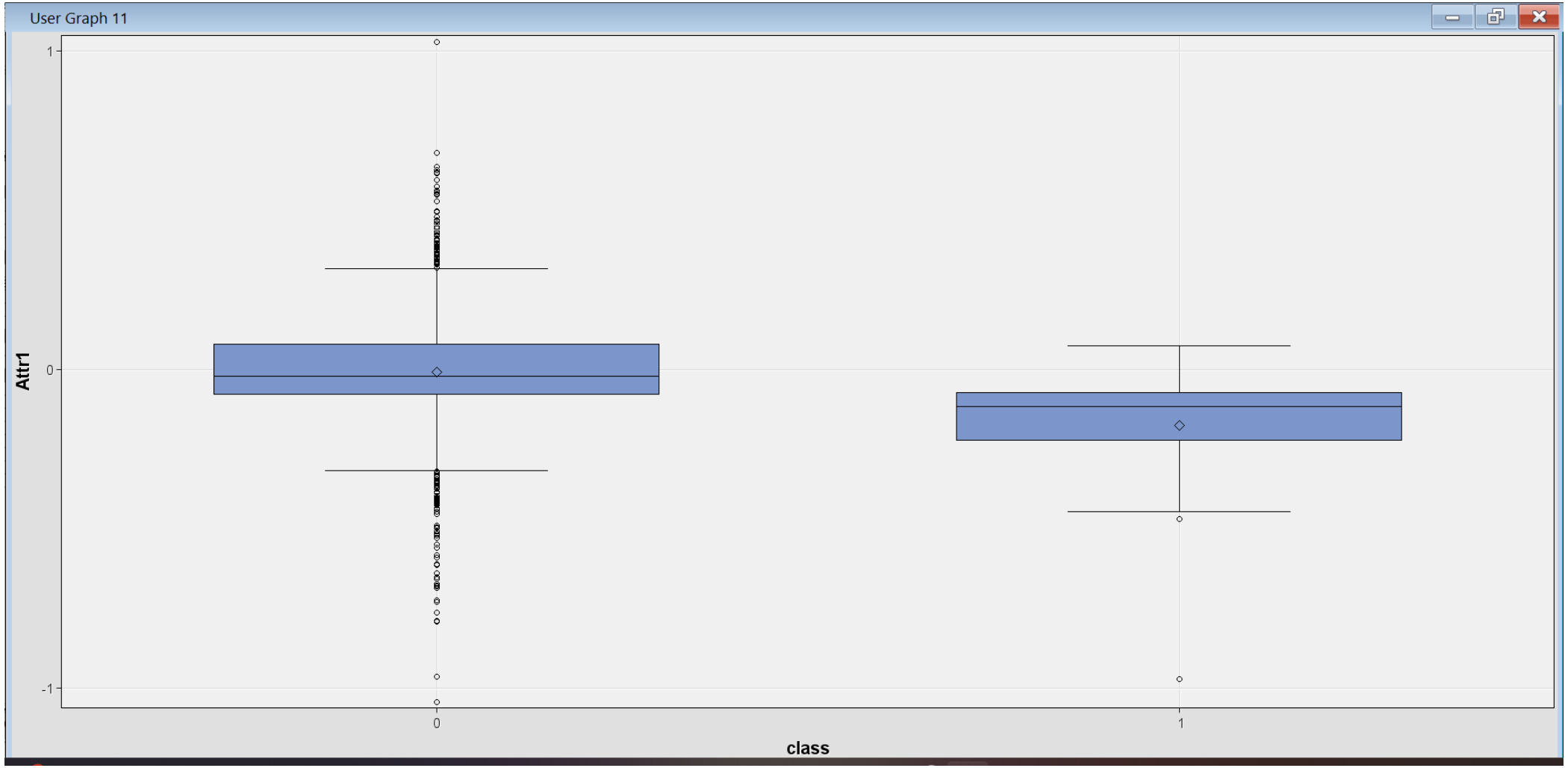
Attribute 1 – After Filter Node



Attribute 1 – Before Filter Node

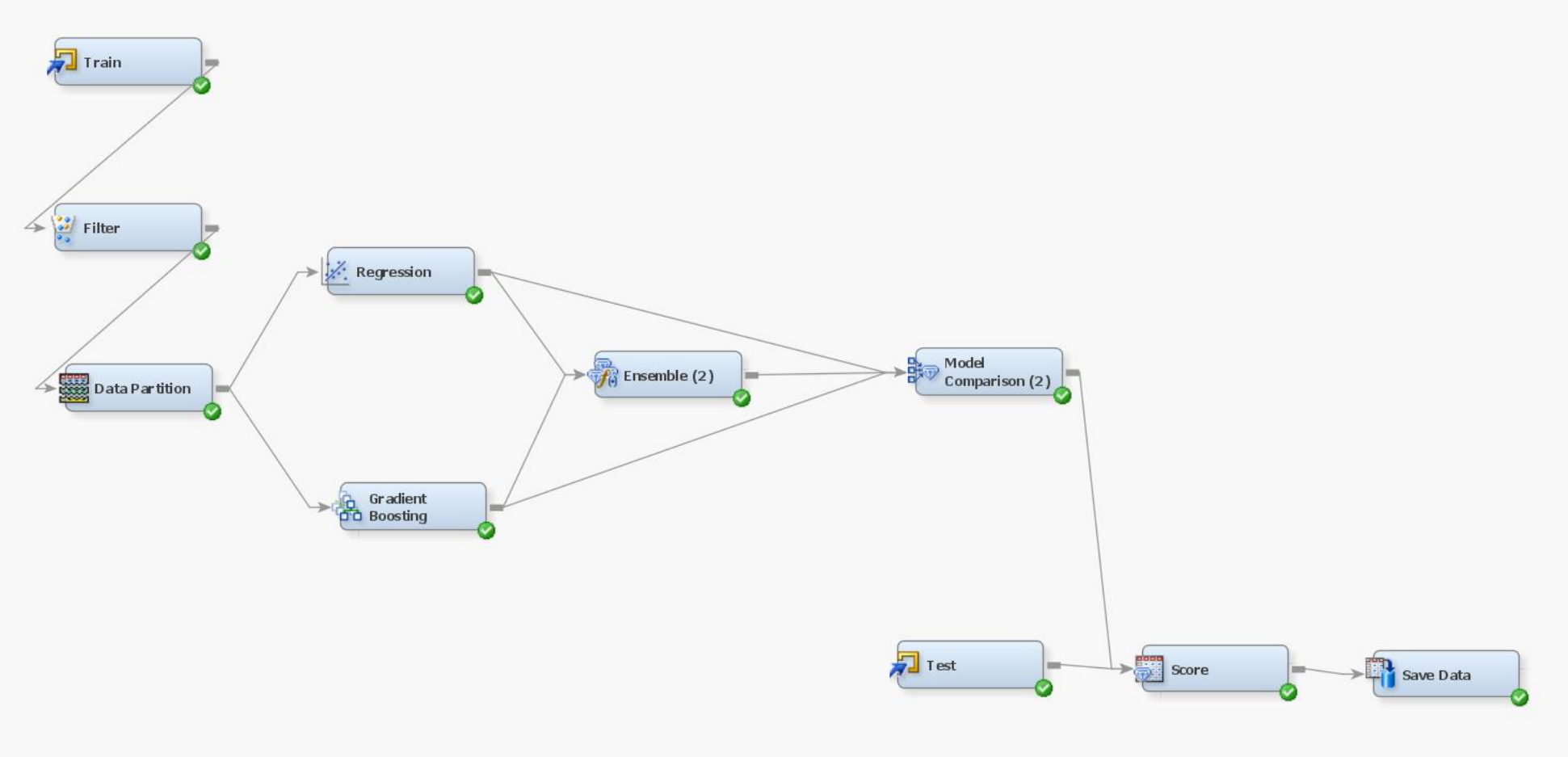


Attribute 1 – After Filter Node



Model 1

Ensemble model for Regression and Gradient Boosting



Data Preprocessing

Filter and Data Partition Nodes

Filter Node

.. Property	Value
General	
Node ID	Filter
Imported Data	
Exported Data	
Notes	
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
Class Variables	
Class Variables	
Default Filtering Method	None
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percentage	0.01
Maximum Number of Levels Cutoff	25
Interval Variables	
Interval Variables	
Default Filtering Method	Standard Deviations from the Mean
Keep Missing Values	Yes
Tuning Parameters	
Score	
Create Score Code	Yes
Update Measurement Level	No
Status	
Create Time	11/26/23 4:44 PM
Run ID	d0679573-98fc-4dc2-b618-45db74da69d
Last Error	
Last Status	Complete
Last Run Time	11/26/23 5:28 PM
Run Duration	0 Hr. 0 Min. 26.81 Sec.
Grid Host	
User-Added Node	No

Data Partition Node

.. Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	54321
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	11/26/23 4:44 PM
Run ID	653b49eb-0486-474d-b44a-0ffdc2928b2a
Last Error	
Last Status	Complete
Last Run Time	11/28/23 7:53 PM
Run Duration	0 Hr. 0 Min. 3.59 Sec.
Grid Host	
User-Added Node	No

Updated Training data to 70% and Validation data to 30%

PURDUE UNIVERSITY

Model

Basic model configurations

Regression Node

.. Property	Value
General	
Node ID	Reg
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Y
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic 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
Convergence Criteria	
Uses Defaults	Yes
Options	
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	No

- *Polynomial Regression with 2 Degrees*
- *Logistic Regression because it's a classification problem*

Gradient Boosting Node

.. Property	Value
General	
Node ID	Boost
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.05
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Decision
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes
Report	
Observation Based Importance	No
Number Single Var Importance	5

- *50 Iterations*
- *0.05 learning rate*
- *Max Tree Depth is 10*
- *Proportion of Data that we train on is 60%*

Results: Model 1

Ensemble Node Results

Classification Results

Classification Table

Data Role=TRAIN Target Variable=class Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	99.591	100.000	6090	97.8942
1	0	0.409	19.084	25	0.4019
1	1	100.000	80.916	106	1.7039

Data Role=VALIDATE Target Variable=class Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	98.5606	99.6553	2602	97.5262
1	0	1.4394	66.6667	38	1.4243
0	1	32.1429	0.3447	9	0.3373
1	1	67.8571	33.3333	19	0.7121

Event Classification Table

Data Role=TRAIN Target=class Target Label=' '

False Negative	True Negative	False Positive	True Positive
25	6090	.	106

Data Role=VALIDATE Target=class Target Label=' '

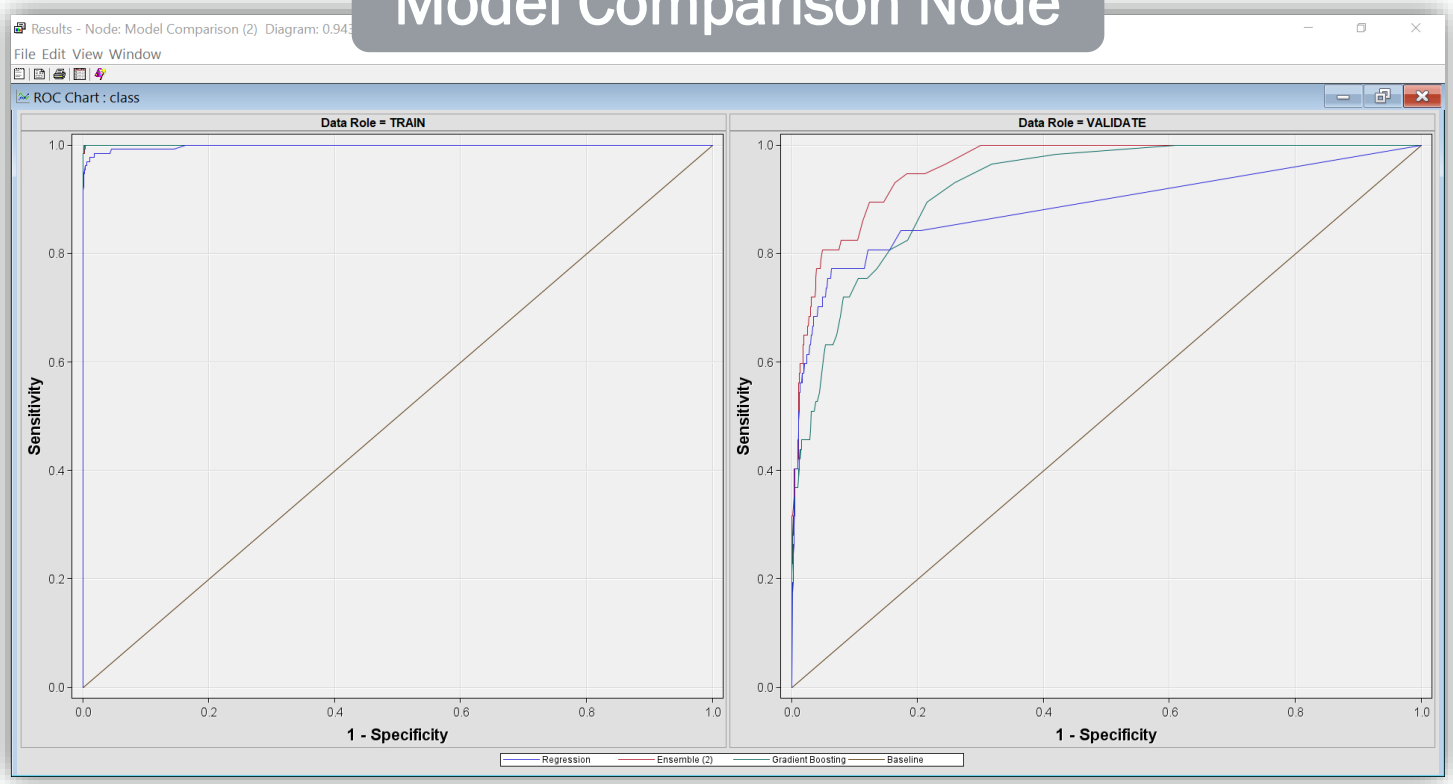
False Negative	True Negative	False Positive	True Positive
38	2602	9	19

PURDUE UNIVERSITY

Results: Model 1

Model Comparison Node and Kaggle Score

Model Comparison Node

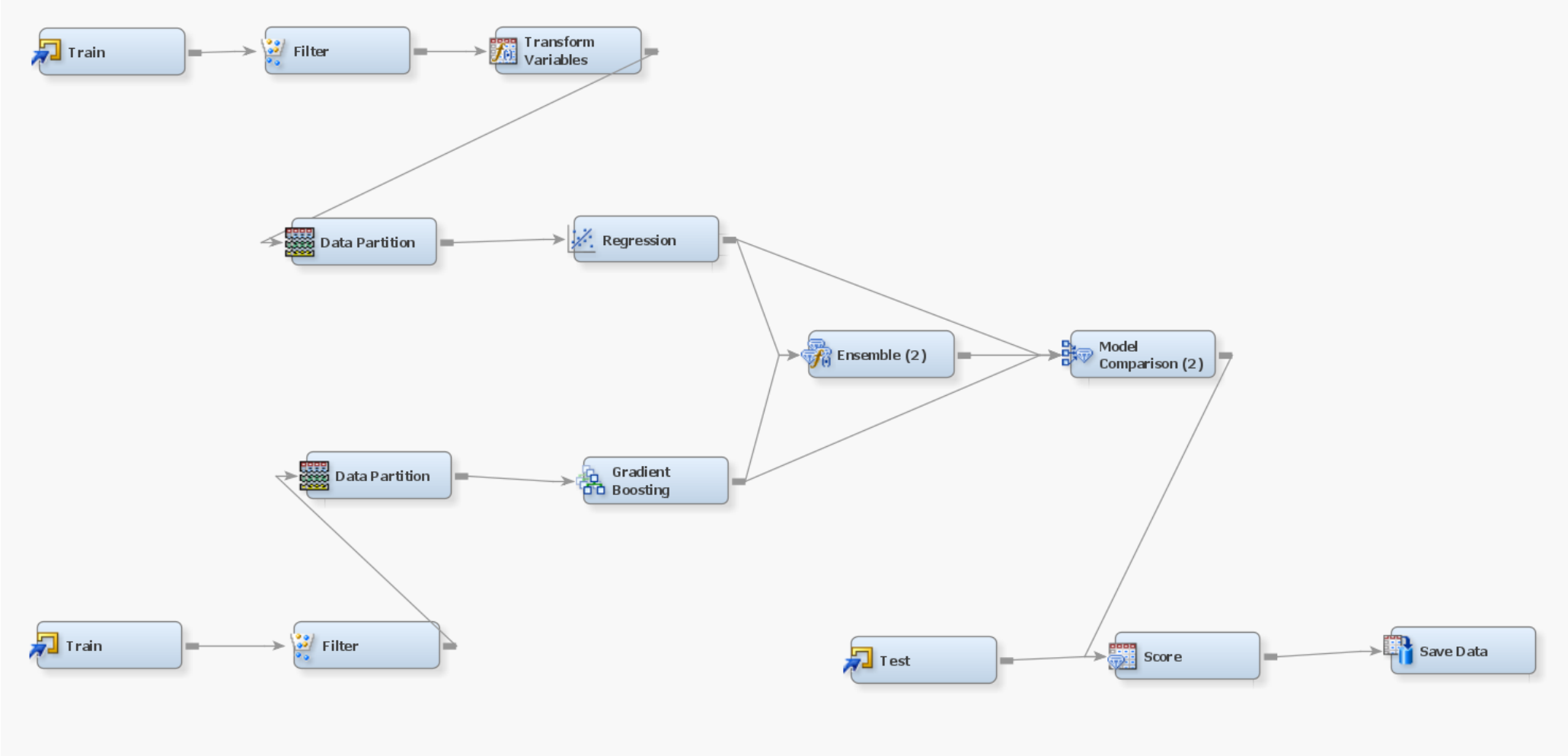


- Ensemble model has the **best ROC** maximizing the AUC
- **Best Model ROC:** 0.959
- **Public Leaderboard Score:** 0.9433
- **Private Leaderboard Score:** 0.94133

PURDUE UNIVERSITY

Model 2

Ensemble model for Regression and Gradient Boosting with Transform Variables



Model 2

Extra Layer of Data Preprocessing

Transform Variables Node

Train	
Variables	
Formulas	
Interactions	
SAS Code	
Default Methods	
Interval Inputs	Log
Interval Targets	None
Class Inputs	None
Class Targets	None
Treat Missing as Level	No
Sample Properties	
Method	First N
Size	Default
Random Seed	12345
Optimal Binning	
Number of Bins	4
Missing Values	Use in Search
Grouping Method	
Cutoff Value	0.1
Group Missing	No
Number of Bins	Variables
Add Minimum Value to Offset Value	Yes
Offset Value	1
Score	
Use Meta Transformation	Yes
Hide	Yes
Reject	Yes
Report	
Summary Statistics	Yes
Status	
Create Time	11/28/23 8:23 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Log transformation on interval variables:
Log transformation is a crucial step for:

- Stabilizing variance
- Normalizing distributions
- Making the data more suitable for analysis in regression models.

PURDUE UNIVERSITY

Results: Model 2

Ensemble Node Results

Classification Results

Classification Table

Data Role=TRAIN Target Variable=class Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	99.089	100.000	6090	97.8942
1	0	0.911	42.748	56	0.9002
1	1	100.000	57.252	75	1.2056

Data Role=VALIDATE Target Variable=class Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	98.3786	99.9234	2609	97.7886
1	0	1.6214	75.4386	43	1.6117
0	1	12.5000	0.0766	2	0.0750
1	1	87.5000	24.5614	14	0.5247

n Table

Data Role=TRAIN Target=class Target Label=' '

False Negative	True Negative	False Positive	True Positive
56	6090	.	75

Data Role=VALIDATE Target=class Target Label=' '

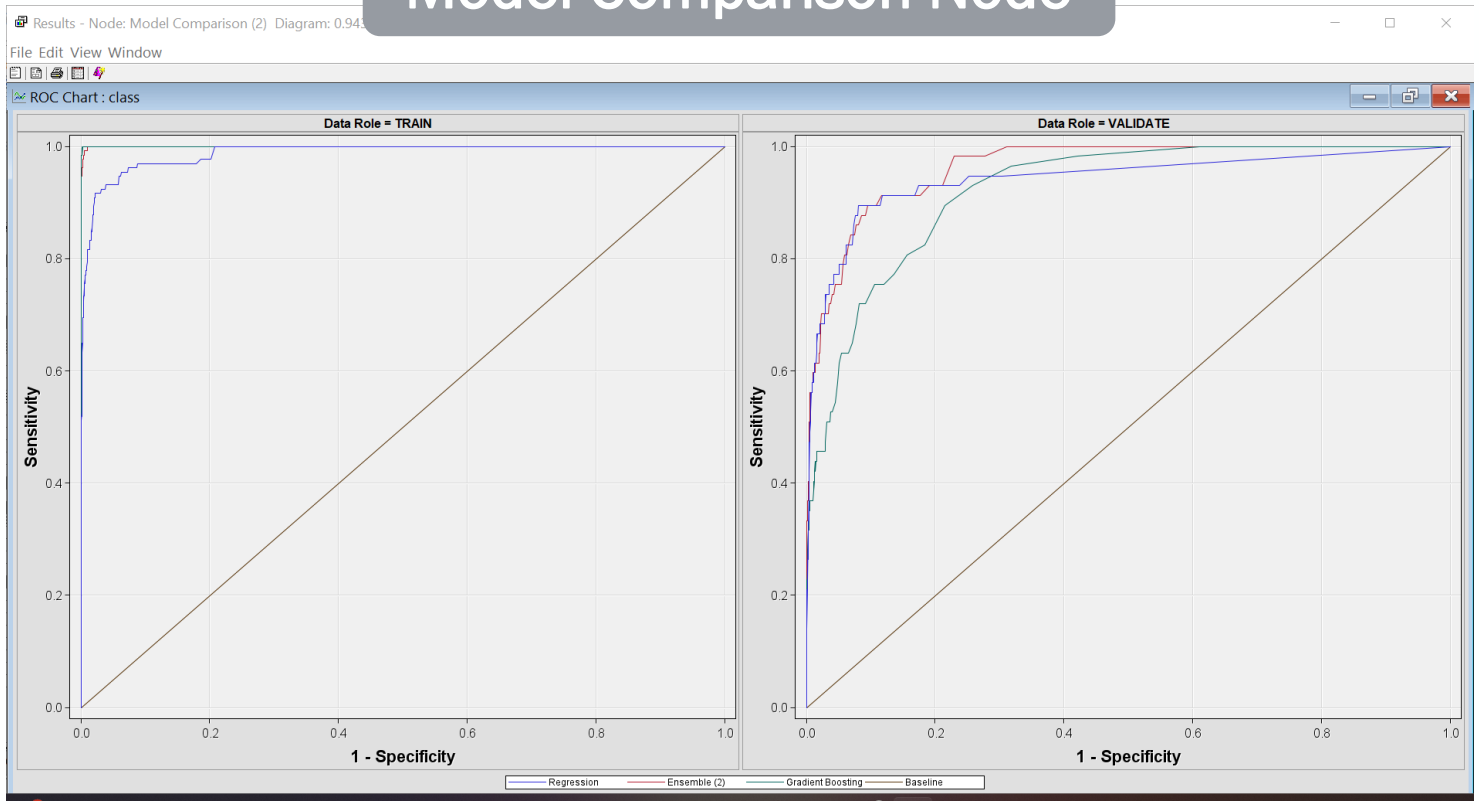
False Negative	True Negative	False Positive	True Positive
43	2609	2	14

PURDUE UNIVERSITY

Results: Model 2

Model Comparison Node and Kaggle Score

Model Comparison Node



- Ensemble model has the **best ROC** maximizing the AUC
- **Best Model ROC:** 0.965
- **Public Leaderboard Score:** 0.94319
- **Private Leaderboard Score:** 0.95262

PURDUE UNIVERSITY

Other Models

Models Evaluated and Rejected during the Analysis

Random Forest: Overfitting and poor performance on validation set.

Neural Network

Different Combinations of the models in the Ensemble Node.

PURDUE UNIVERSITY

Thank You

PURDUE UNIVERSITY®