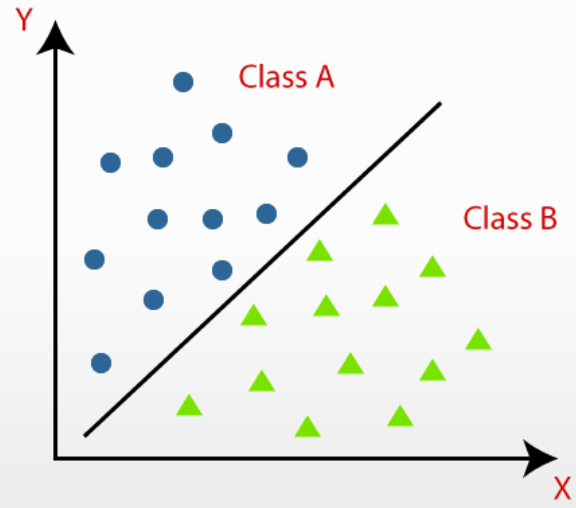
Bankruptcy Prediction

MGMT 571: Final Project

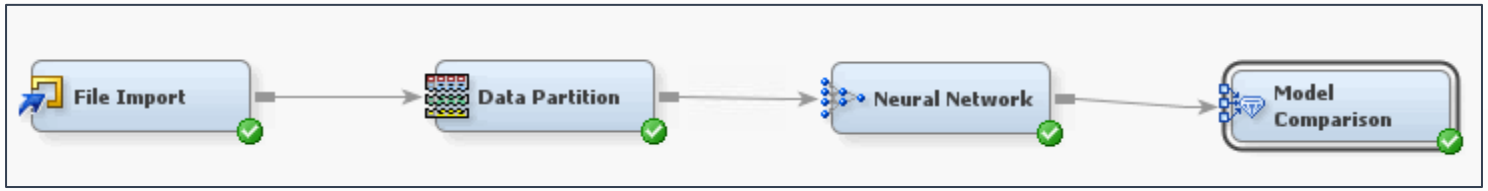
Team 512MB
Archita Ray
Rahul Chowdary Kunku
Sai Mona Duvvapu

About the Dataset

- 64 Independent Variables
- 1 Dependent Variable
- 1 Training Dataset
- 1 Test Dataset
- Output: **EventProbability**
- The estimated probability that an observation or data point belongs to the positive class.



Initial Approach: Neural Network



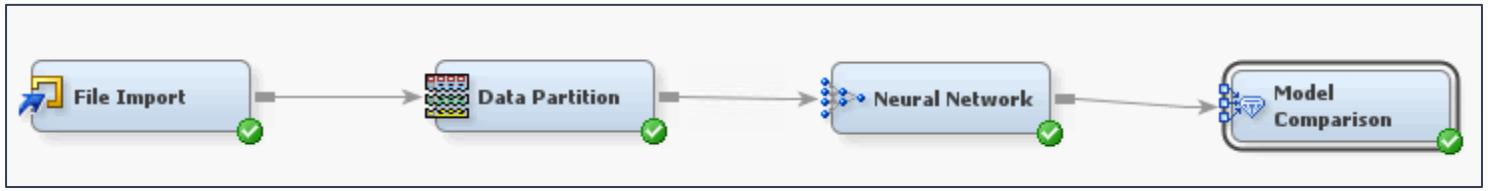
Data Partition

General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	36402784
Data Set Allocations	
Training	75.0
Validation	25.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	

Results

Model Description	Selection Criterion: Train: Average Squared Error ▼	Train: Roc Index	Valid: Roc Index
Neural Network	0.012685	0.954	0.905

Initial Approach: Neural Network



Property	Value
General	
Node ID	Neural
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	36402784
Model Selection Criterion	Average Error
Suppress Output	No
Score	
Hidden Units	Yes
Residuals	Yes
Standardization	No

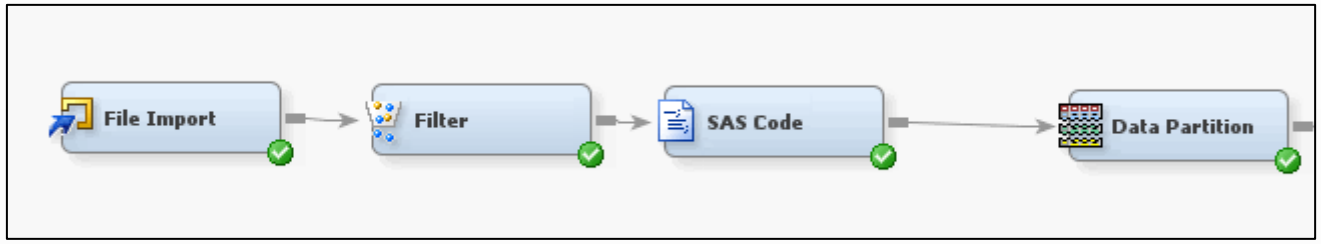
Network

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 F	Default
Hidden Layer Activation Fun	Default
Hidden Bias	Yes
Target Layer Combination F	Default
Target Layer Activation Fun	Default

Optimization

Property	Value
Training Technique	Default
Maximum Iterations	50
Maximum Time	4 Hours
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

Learnings along the way: Preprocessing

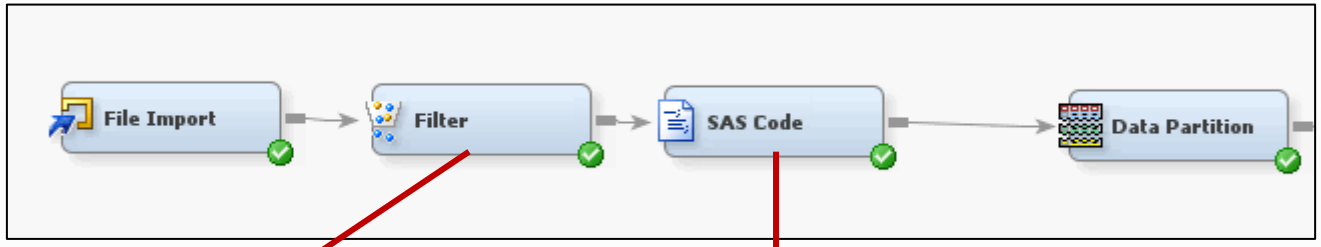


Results

Model Description	Selection Criterion: Train: Average Squared Error	Train: Roc Index	Valid: Roc Index
Neural Network	0.010199	0.971	0.927
Gradient Boosting	0.013961	0.951	0.901
Regression (2)	0.014305	0.927	0.935

- Filtering out data that's 3 standard deviations away from the mean for all variables did well with Train data
- However, the model failed to get a reasonable AUC for the public test data

Learnings along the way: Preprocessing



Train	
Export Table	Filtered
Tables to Filter	All Data Sets
Distribution Data Sets	Yes
Class Variables	
Class Variables	
Default Filtering Method	Rare Values (Percentage)
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percenta	0.01
Maximum Number of Levels	25
Interval Variables	
Interval Variables	
Default Filtering Method	Standard Deviations from the
Keep Missing Values	No
Tuning Parameters	
Score	
Create Score Code	Yes
Update Measurement Level	No

Property	Value
General	
Node ID	EMCODE
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Code Editor	
Tool Type	Modify
Data Needed	Yes
Rerun	No
Use Priors	Yes
Score	
Advisor Type	Basic
Publish Code	Publish
Code Format	DATA step

Training Code

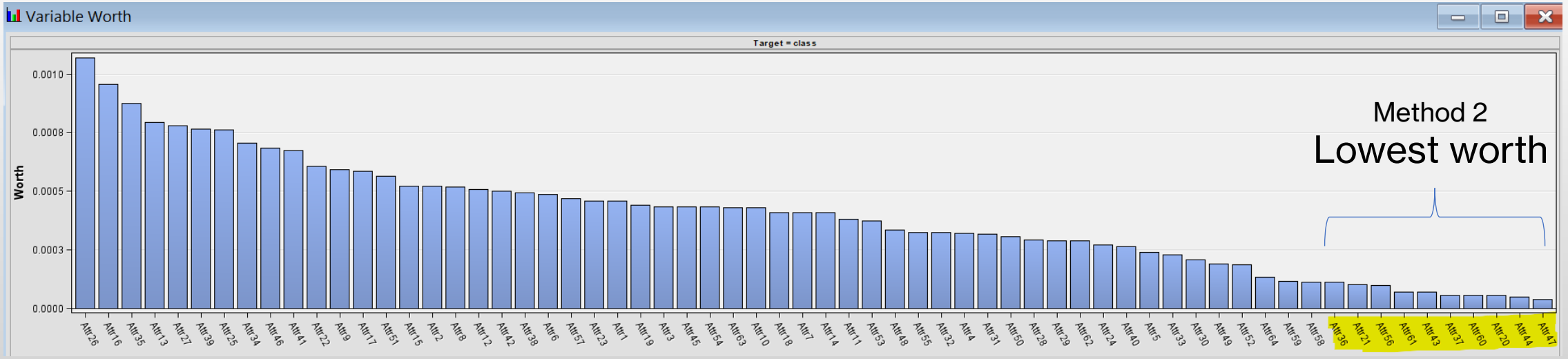
```
PROC SQL;
   CREATE TABLE clean_data AS
   SELECT DISTINCT *
   FROM EMWS5.FIMPORT_DATA;
```

Learnings along the way: Preprocessing

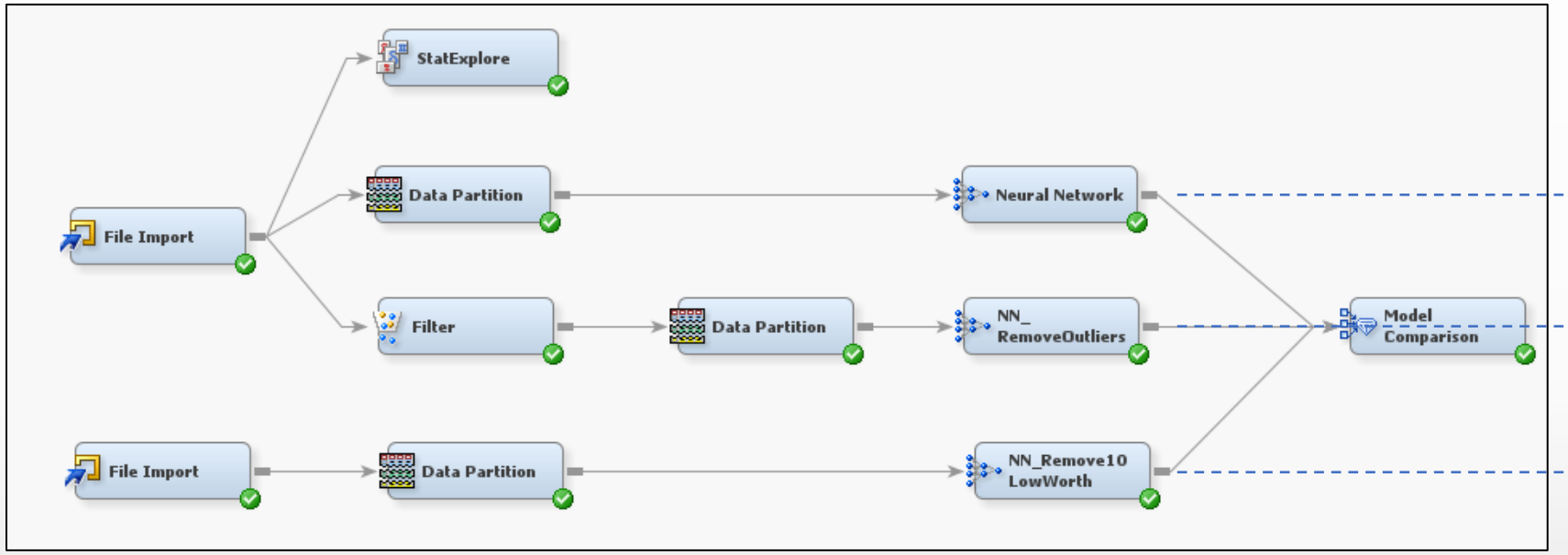
Name	Report	Filtering Method
Attr18	No	Extreme Percentiles
Attr14	No	Extreme Percentiles
Attr1	No	Extreme Percentiles
Attr62	No	Extreme Percentiles
Attr7	No	Extreme Percentiles
Attr2	No	Extreme Percentiles
Attr6	No	Extreme Percentiles
Attr38	No	Extreme Percentiles
Attr10	No	Extreme Percentiles
Attr39	No	Extreme Percentiles

Method 1

Filtered variables due to high kurtosis and skewness



Learnings along the way: Preprocessing



Changes in Data Partition

Filtering extreme percentile for 10 variables

Rejecting 10 variables with lowest worth

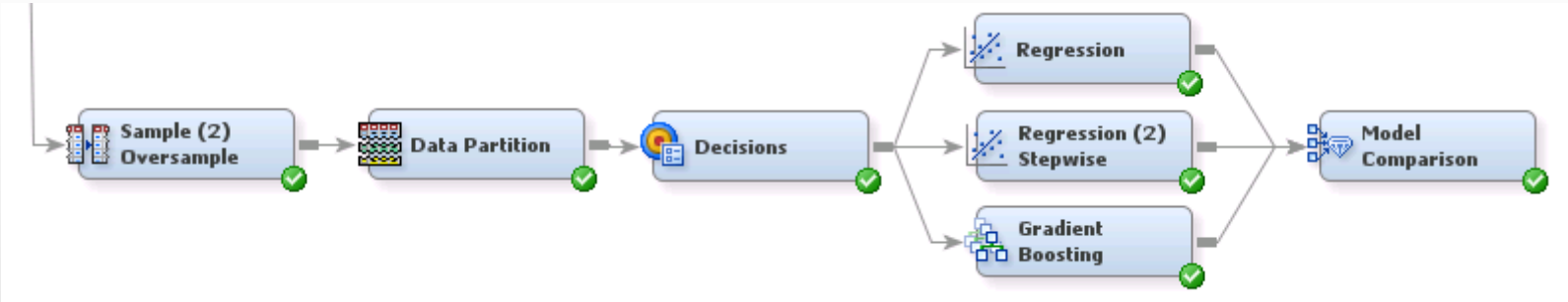
Results

Model Description	Selection Criterion: Train: Roc Index	Train: Average Squared Error	Valid: Average Squared Error	Valid: Roc Index
NN_Remove10LowWorth	0.939	0.015623	0.020184	0.885
NN_RemoveOutliers	0.931	0.015354	0.016636	0.884
Neural Network	0.926	0.016642	0.019044	0.881

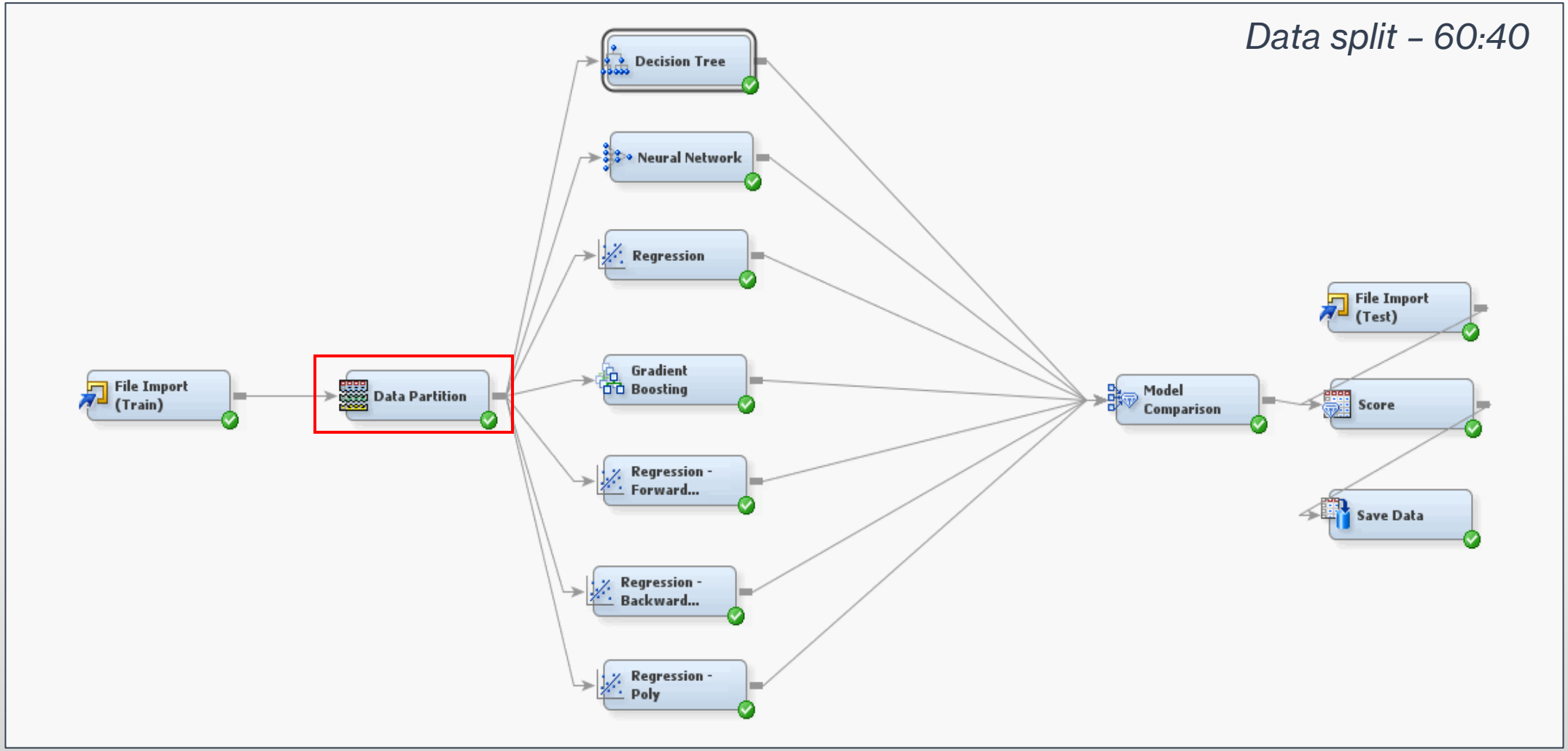
- No significant effect on valid ROC
- Submissions based on better train ROC did not do any better with public test data

Learnings along the way: Oversampling

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	class	TARGET	0	6851	97.8854
TRAIN	class	TARGET	1	148	2.1146

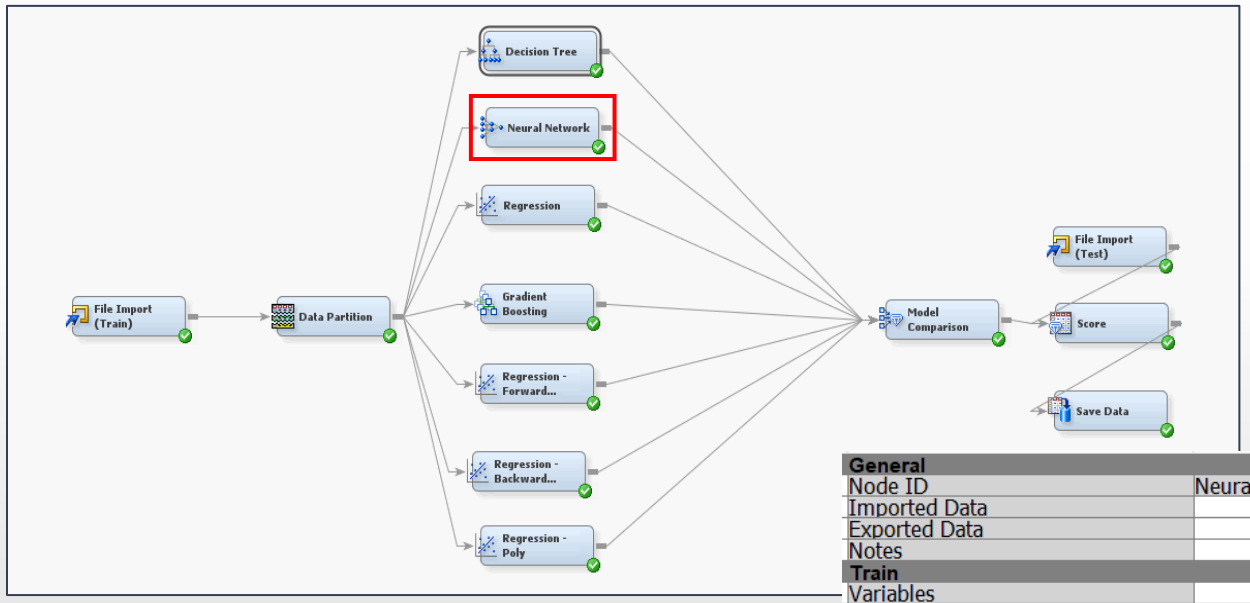


Our second-best model



Our second-best model

Neural Network



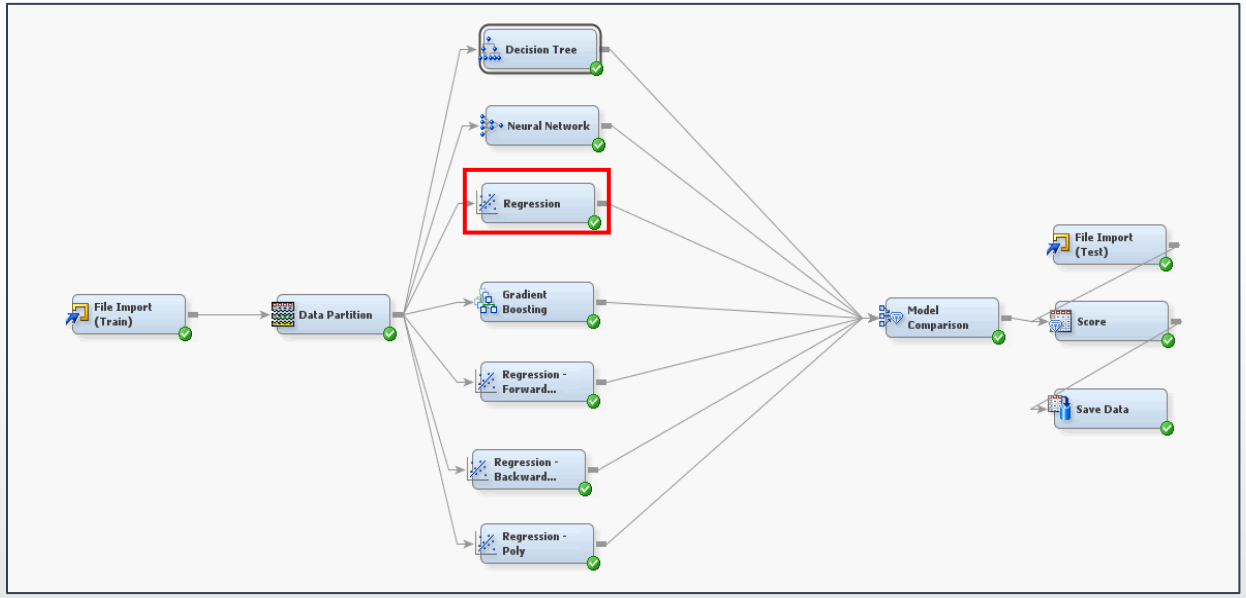
General	
Node ID	Neural
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Continue Training	Yes
Network	
Optimization	
Initialization Seed	36723317
Model Selection Criterion	Average Error
Suppress Output	No
Score	
Hidden Units	Yes
Residuals	Yes
Standardization	No
Status	

Optimization

Property	Value
Training Technique	RProp
Maximum Iterations	50
Maximum Time	4 Hours
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

Our second-best model

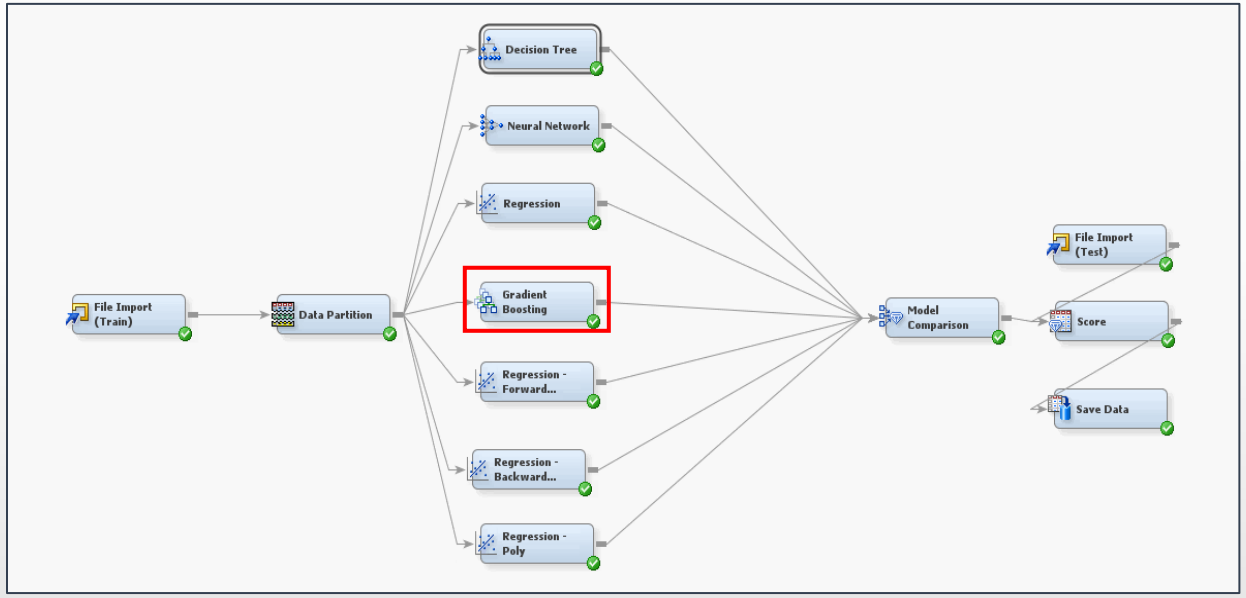
Logistic Regression



Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
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	Yes
Correlation	Yes
Statistics	Yes
Suppress Output	No
Details	Yes
Design Matrix	No

Our second-best model

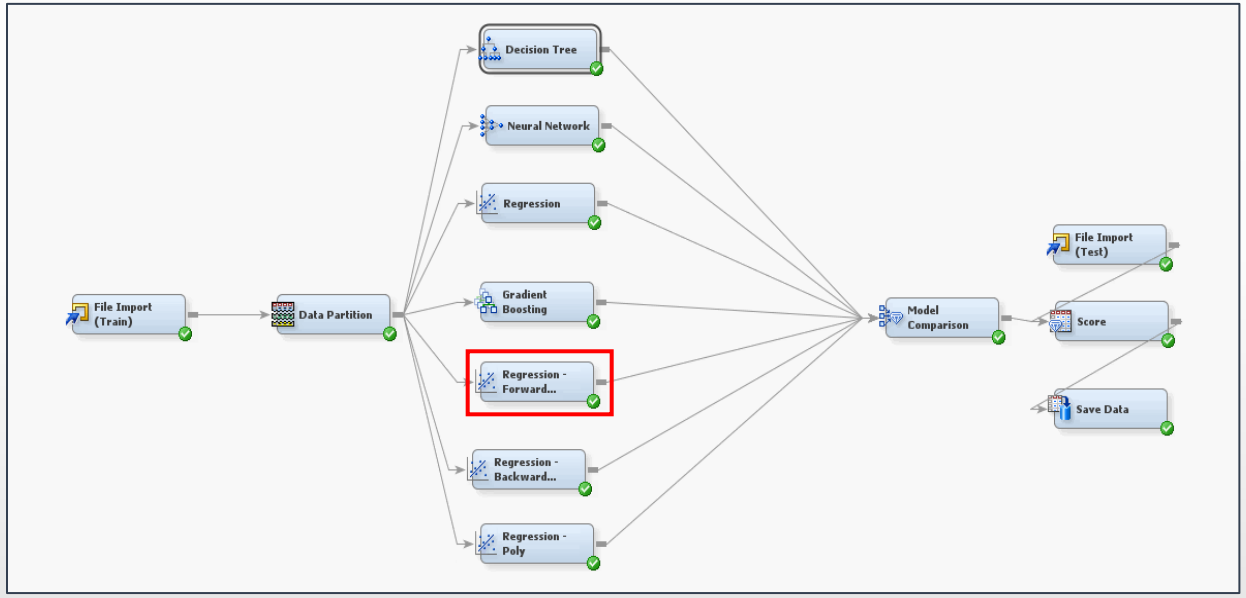
Gradient Boosting



Series Options	
N Iterations	50
Seed	36723317
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
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
Status	

Our second-best model

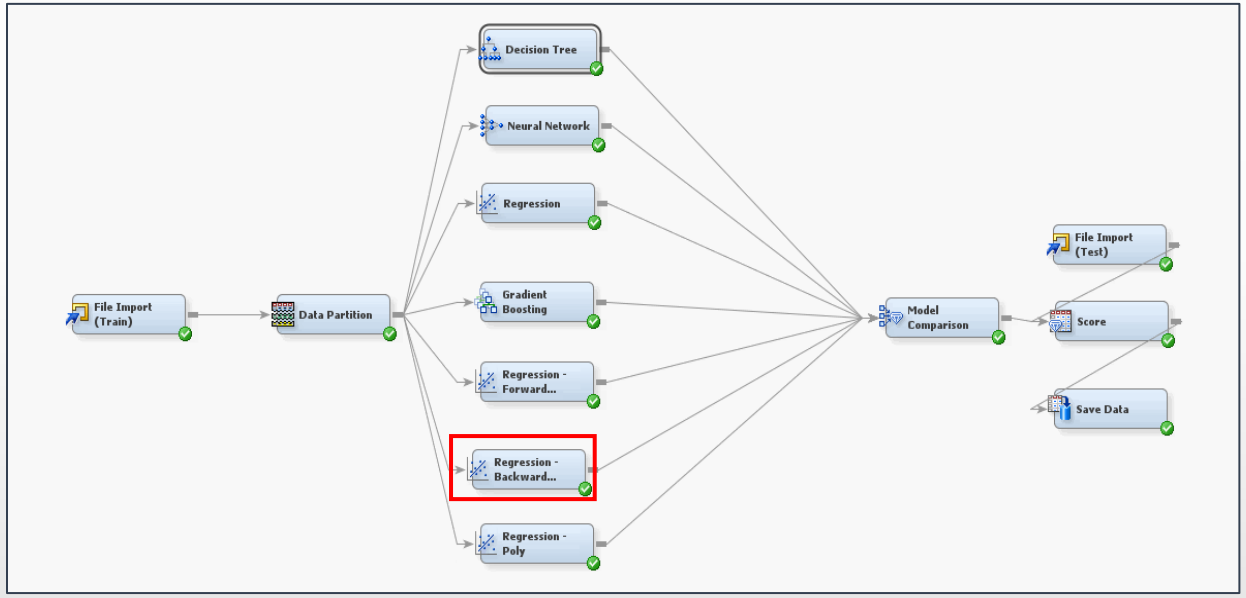
Forward Regression



Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
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	Forward
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	Yes
Correlation	Yes
Statistics	Yes
Suppress Output	No
Details	Yes
Design Matrix	No
Score	
Excluded Variables	Reject

Our second-best model

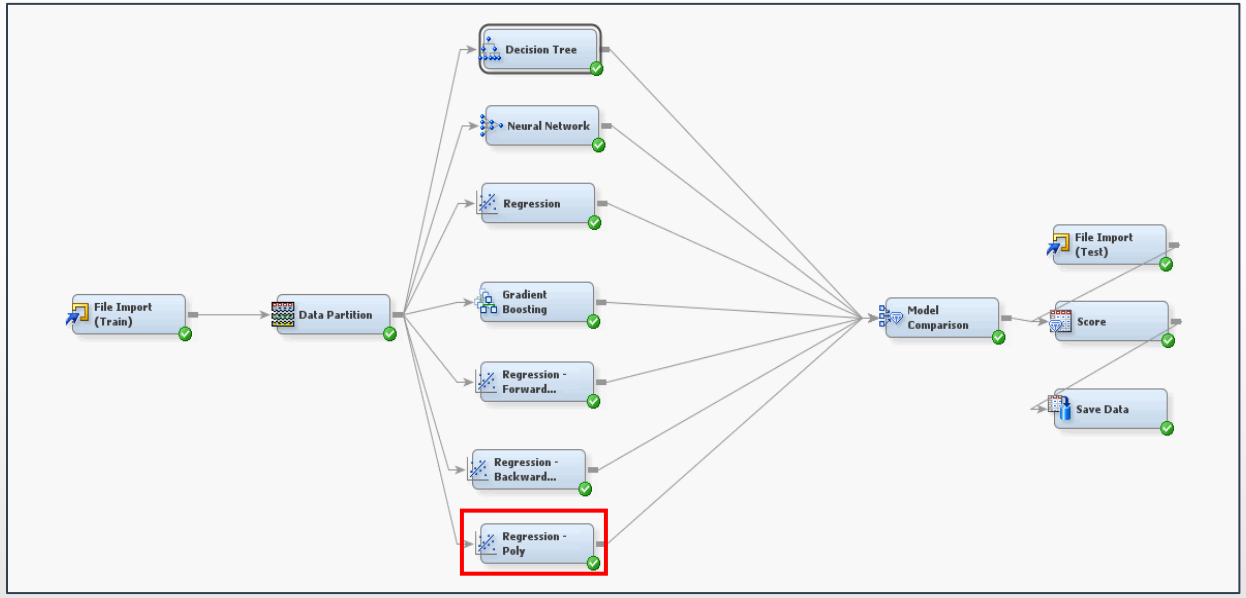
Backward Regression



Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
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	Backward
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	Yes
Correlation	Yes
Statistics	Yes
Suppress Output	No
Details	Yes
Design Matrix	No
Score	
Excluded Variables	Reject

Our second-best model

Polynomial Regression

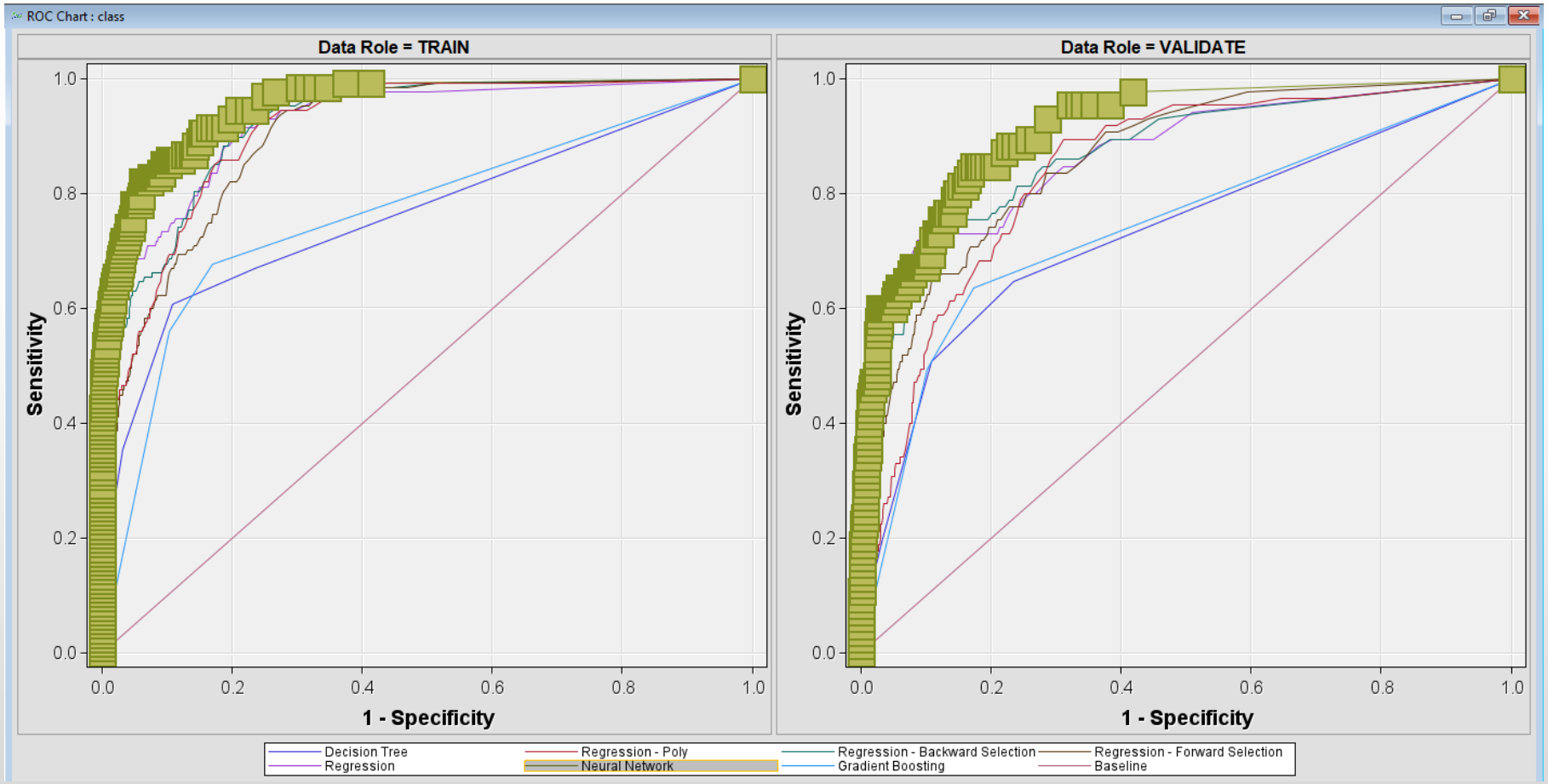


Equation	
Main Effects	Yes
Two-Factor Interactions	Yes
Polynomial Terms	Yes
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection 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	Yes
Correlation	Yes
Statistics	Yes
Suppress Output	No
Details	Yes
Design Matrix	No
Score	
Excluded Variables	Reject

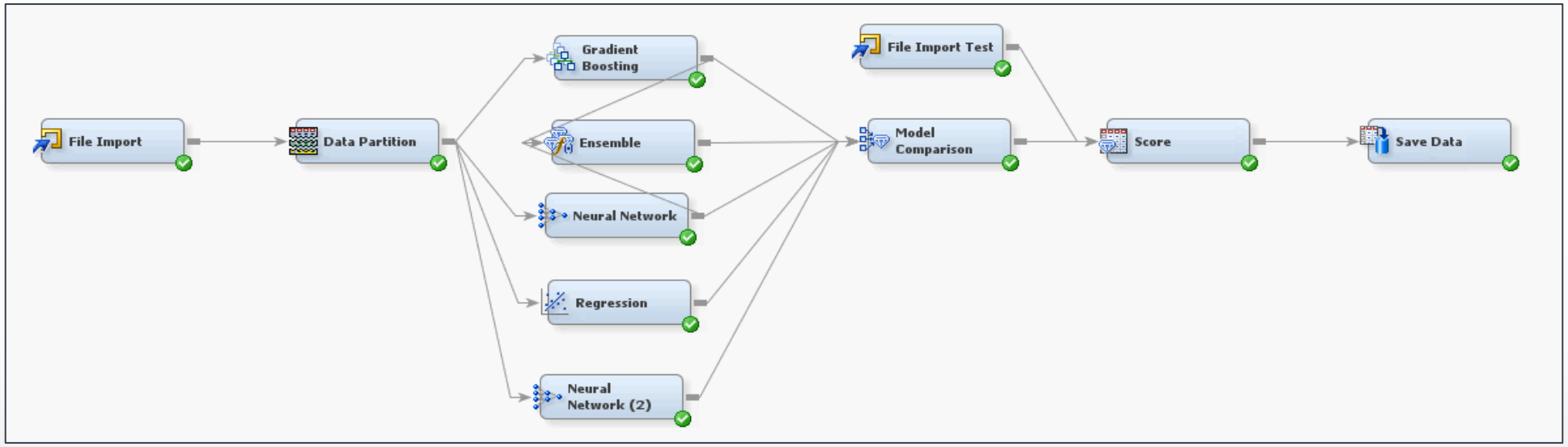
Our second-best model- Results

Model Description	Valid: Roc Index	Train: Roc Index	Train: Average Squared Error	Valid: Average Squared Error	Train: Misclassification Rate	Valid: Misclassification Rate
Neural Network	0.915	0.955	0.011699	0.017739	0.013002	0.020745
Regression	0.869	0.925	0.014156	0.019295	0.01717	0.022744
Regression - Backward Selection	0.873	0.925	0.014653	0.018657	0.01717	0.021245
Regression - Poly	0.837	0.914	0.01695	0.022707	0.019837	0.024494
Regression - Forward Selection	0.859	0.902	0.017282	0.019441	0.020003	0.021495
Decision Tree	0.733	0.763	0.017605	0.019782	0.01867	0.020745
Gradient Boosting	0.777	0.772	0.020471	0.020574	0.02117	0.021245

Our second-best model- Results



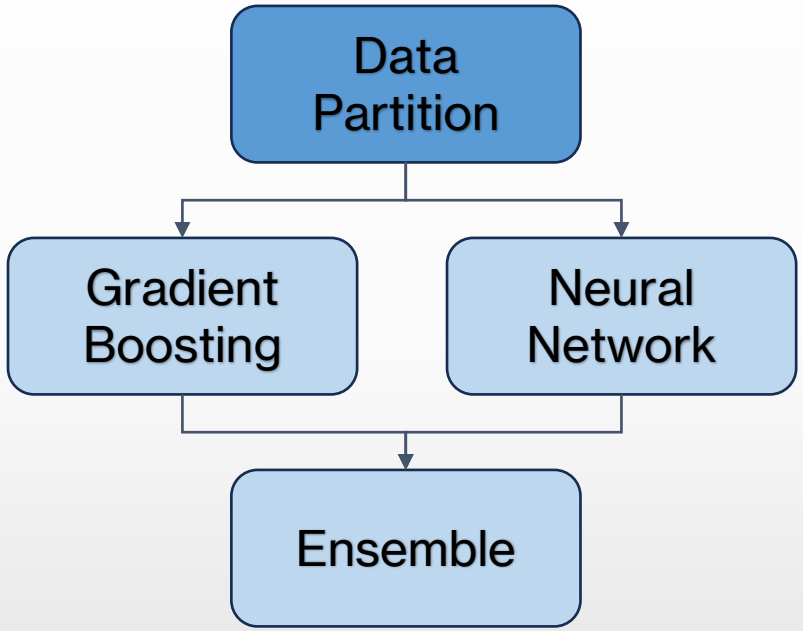
Our best model



Model Comparison Results

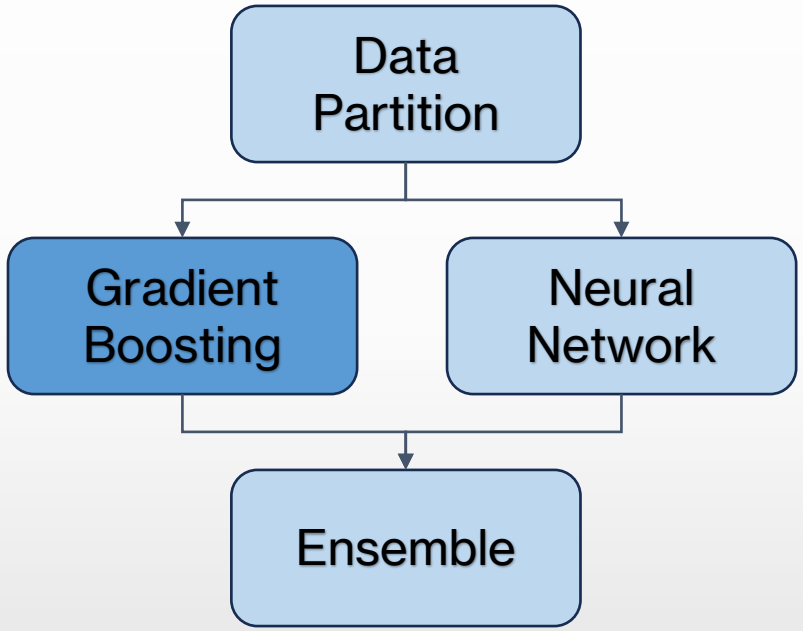
Model Description	Selection Criterion: Valid: Roc Index	Train: Roc Index	Train: Average Squared Error	Valid: Average Squared Error	Train: Misclassification Rate	Valid: Misclassification Rate
Ensemble	0.946	0.992	0.007681	0.014269	0.010269	0.015588
Gradient Boosting	0.918	0.99	0.00854	0.017884	0.01267	0.018385
Neural Network (2)	0.875	0.957	0.013173	0.017066	0.015471	0.019984
Neural Network	0.858	0.98	0.00974	0.01436	0.012403	0.015588
Regression	0.5	0.5	0.020628	0.021117	0.021072	0.021583

Our best model



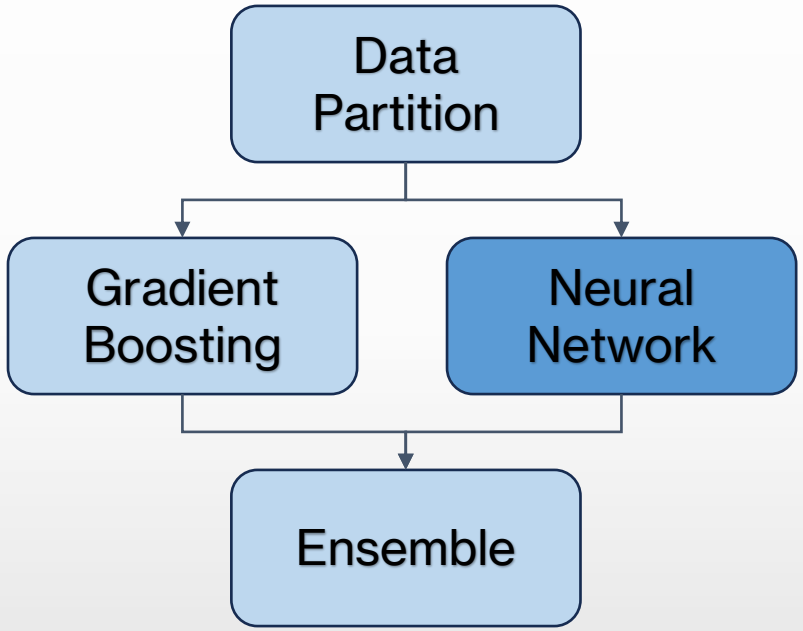
Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	35005605
Data Set Allocations	
Training	75.0
Validation	25.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes

Our best model



Series Options	
N Iterations	50
Seed	35005605
Shrinkage	0.1
Train Proportion	70
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	6
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 Rule	4
Split Size	20
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Misclassification
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes
Report	
Observation Based Import	No
Number Single Var Import	5
Status	

Our best model



General	
Node ID	Neural
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	35005605
Model Selection Criterion	Misclassification
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	
Create Time	11/29/23 5:18 PM
Run ID	ed030b25-b672-4137-91ff-d
Last Error	
Last Status	Complete
Last Run Time	11/29/23 5:53 PM
Run Duration	0 Hr. 0 Min. 9.34 Sec.
Grid Host	
User-Added Node	No

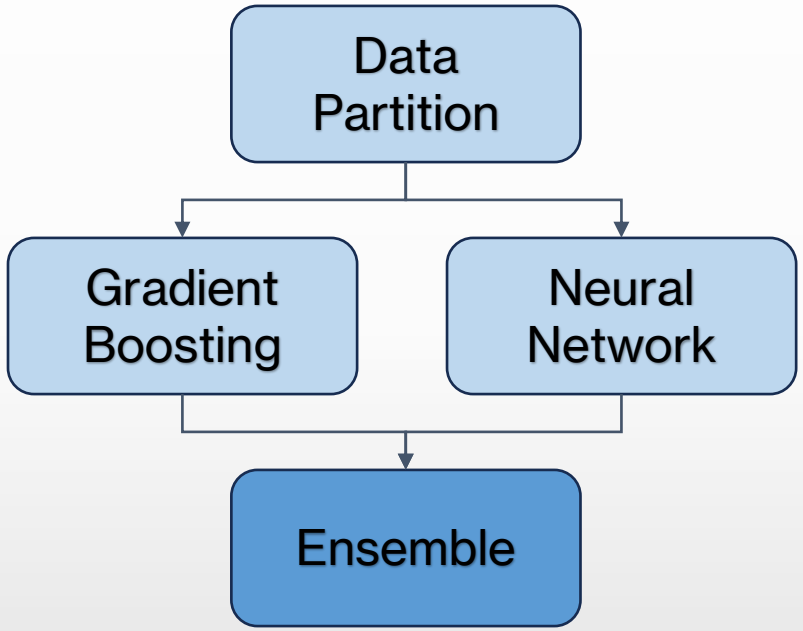
Network

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 Fun	Default
Hidden Layer Activation Functi	Default
Hidden Bias	Yes
Target Layer Combination Func	Default
Target Layer Activation Functio	Default
Target Layer Error Function	Default
Target Bias	Yes
Weight Decay	0.0

Optimization

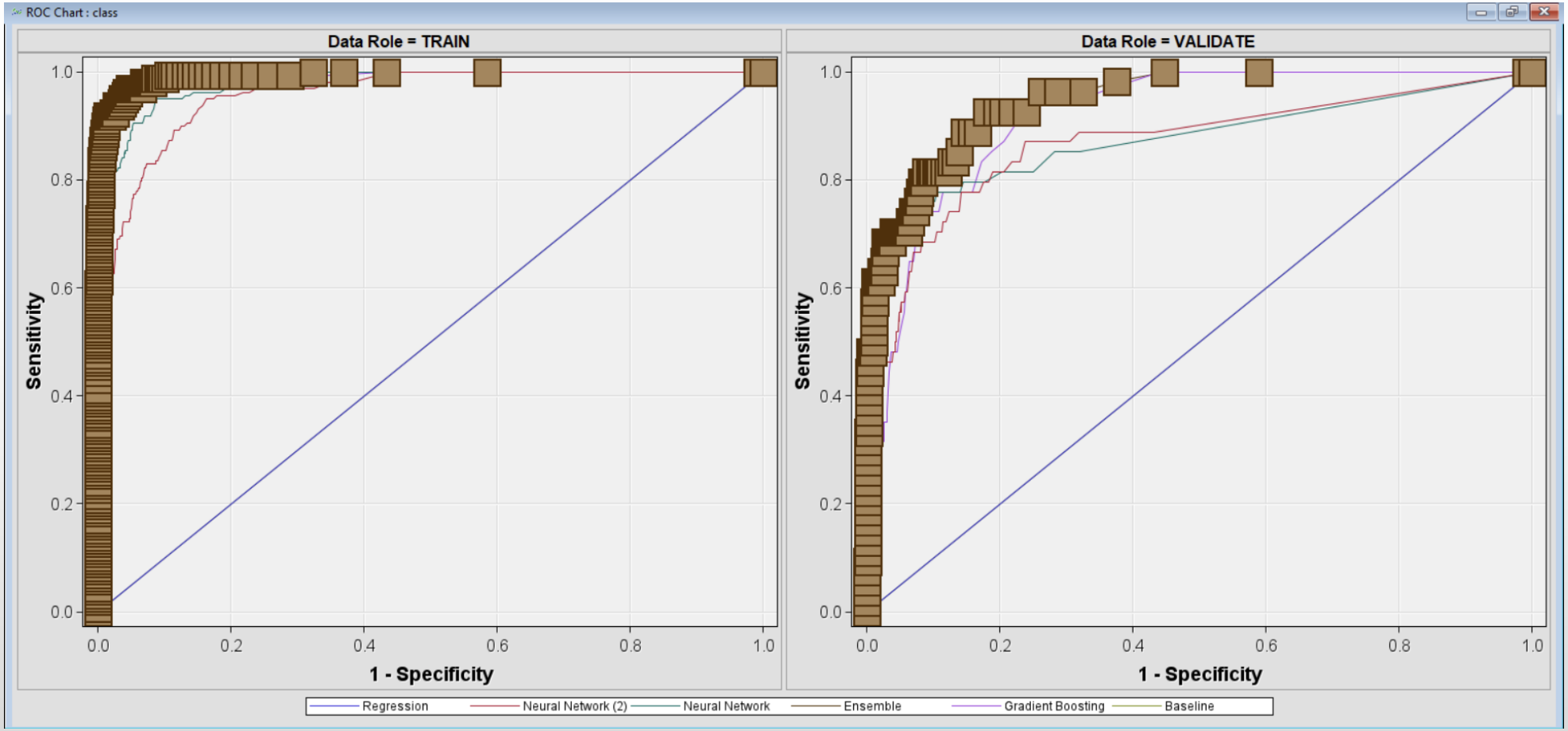
Property	Value
Training Technique	Default
Maximum Iterations	136
Maximum Time	4 Hours
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1
Relative Function	0.0
Relative Function Times	1
Relative Gradient	1.0E-6

Our best model



Property	Value
General	
Node ID	Ensmbl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interval Target	
Predicted Values	Average
Class Target	
Posterior Probabilities	Voting
Voting Posterior Probabilities	Average

Our best model



Takeaways

- Right combination of properties, algorithms is key
- Various metrics need to be evaluated before assuming the goodness of the model
- Finding the balance between a model that's too simple (underfitting) and one that's too complex (overfitting) is critical

THANK YOU