Firm Collapse Prediction

KAGGLE COMPETITION

Team Minecraftsmen

Siddhant Treasure
Soham Agarwal
Varun Annapareddy

The business problem involves stakeholders

01 Allows investors and shareholders to invest money wisely

02 Allows the business to understand which factors are important

03 Enables regulatory bodies to maintain stability in the financial market

04 Customers are affected by the market condition of these large institutions

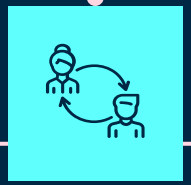
THE TOOL - SAS EM

01

ADVANCED ANALYTICS

Offers sophisticated data mining and predictive modeling

02



CUSTOMIZABLE WORKFLOWS

Enables tailored analytical processes

03

SCALABILITY

Efficiently handles large datasets

SEMMA METHODOLOGY

Step 1

EDA

Step 2

MODIFY

Step 3

MODELLING

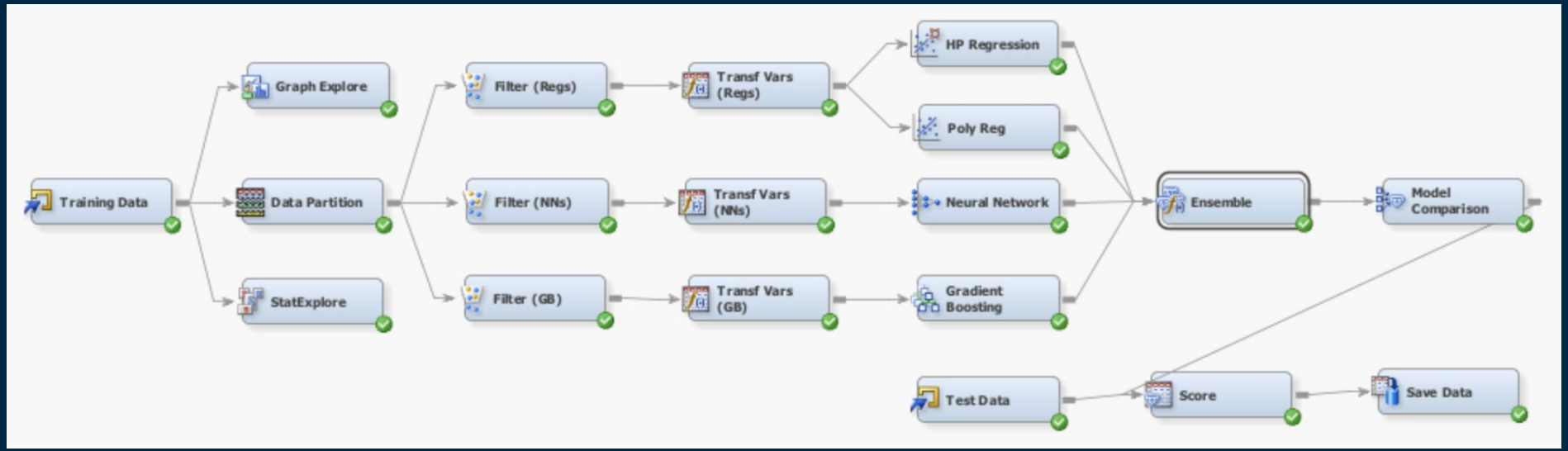
Step 4

ASSESS

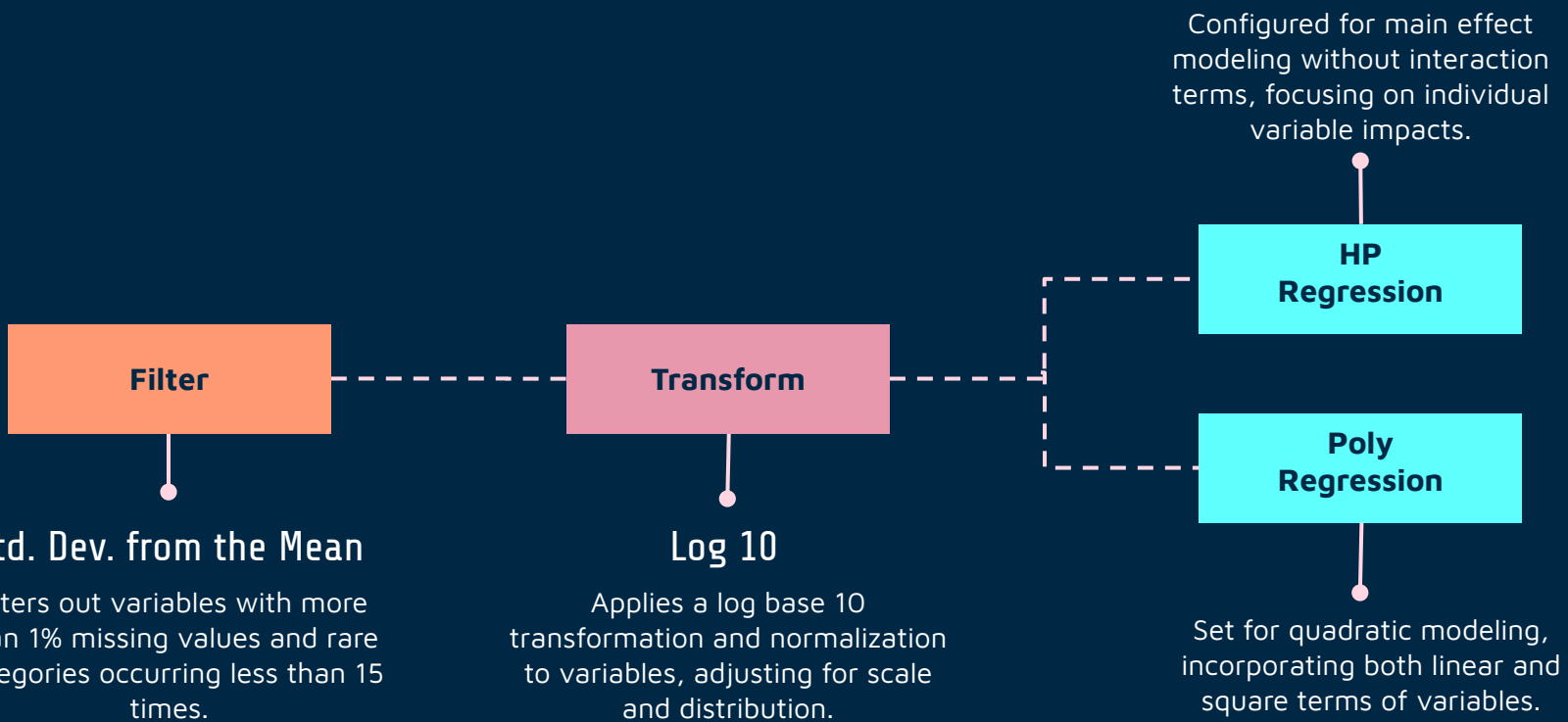
Step 5

A/B Testing
& Hyperparameter tuning

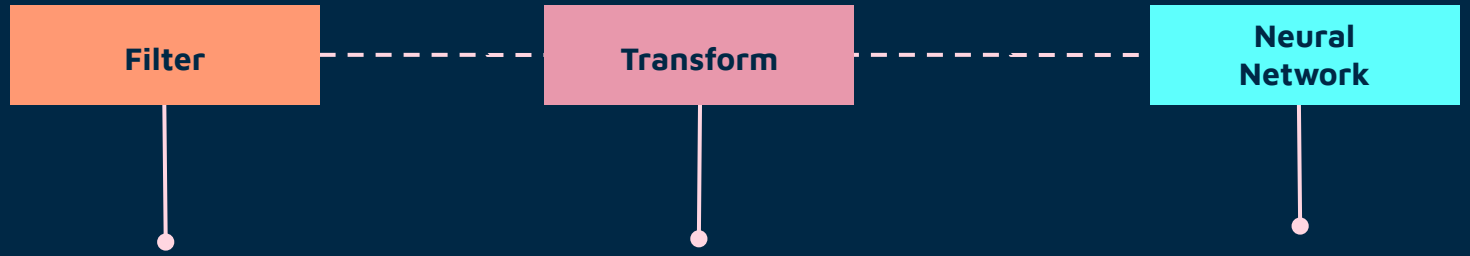
FINAL MODEL – SAS EM



DATA FLOW FOR REGRESSION NODES



DATA FLOW FOR NEURAL NETWORK NODE



Std. Dev. from the Mean

Filters out variables with more than 1% missing values and rare categories occurring less than 15 times.

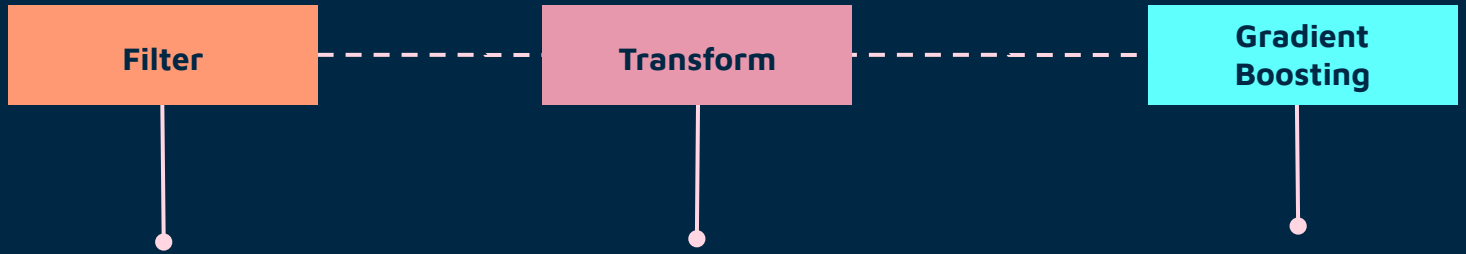
Log 10

Applies a log base 10 transformation and normalization to variables, adjusting for scale and distribution.

Neural Network

Apply complex, non-linear modeling to data for intricate pattern recognition.
4 hidden units and 200 iterations

DATA FLOW FOR GRADIENT BOOSTING NODE

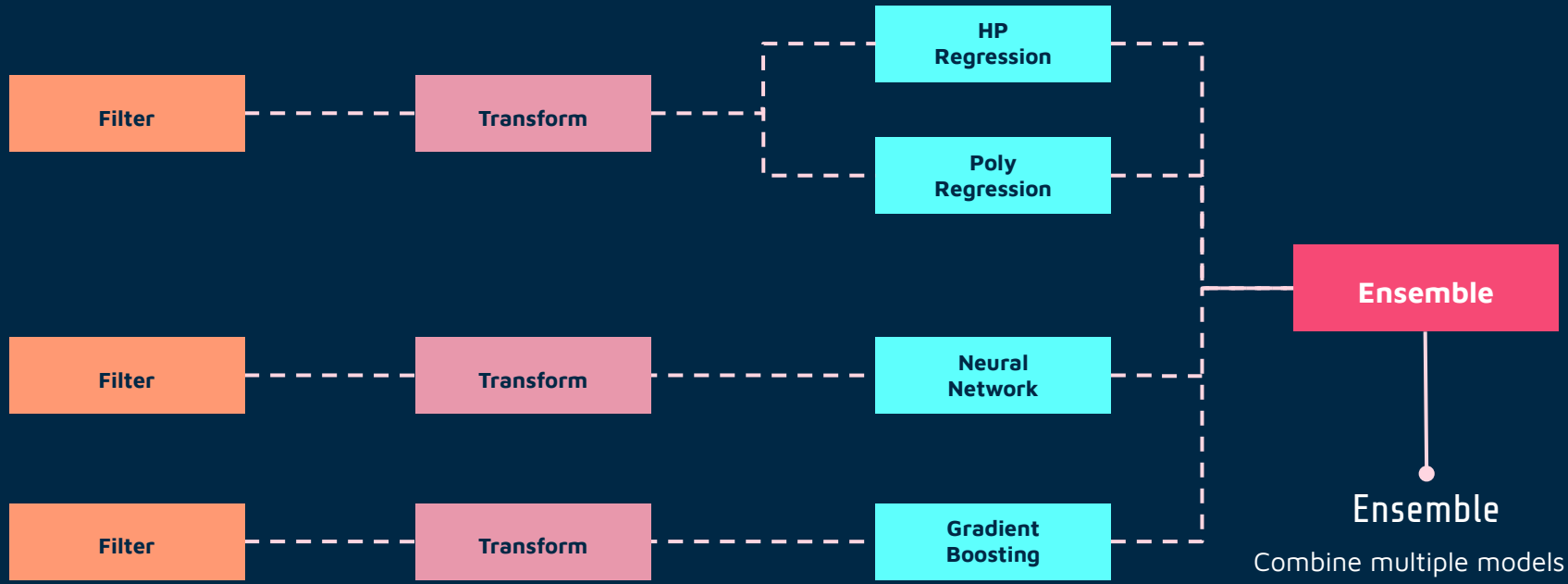


MODEL COMPARISON

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Roc Index
Y	Reg2	Reg2	Poly Reg	class	0.941
	HPReg	HPReg	HP Regression	class	0.931
	Neural	Neural	Neural Network	class	0.915
	Boost	Boost	Gradient Boosting	class	0.9

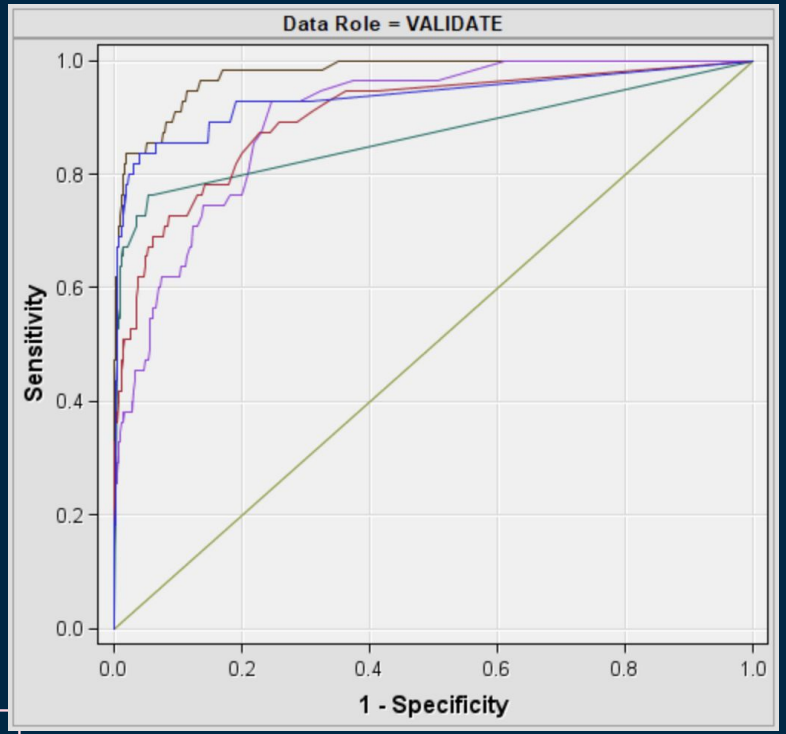
FINAL MODEL



Ensemble

Combine multiple models to improve prediction robustness and accuracy.

MODEL COMPARISON (WITH ENSEMBLE)



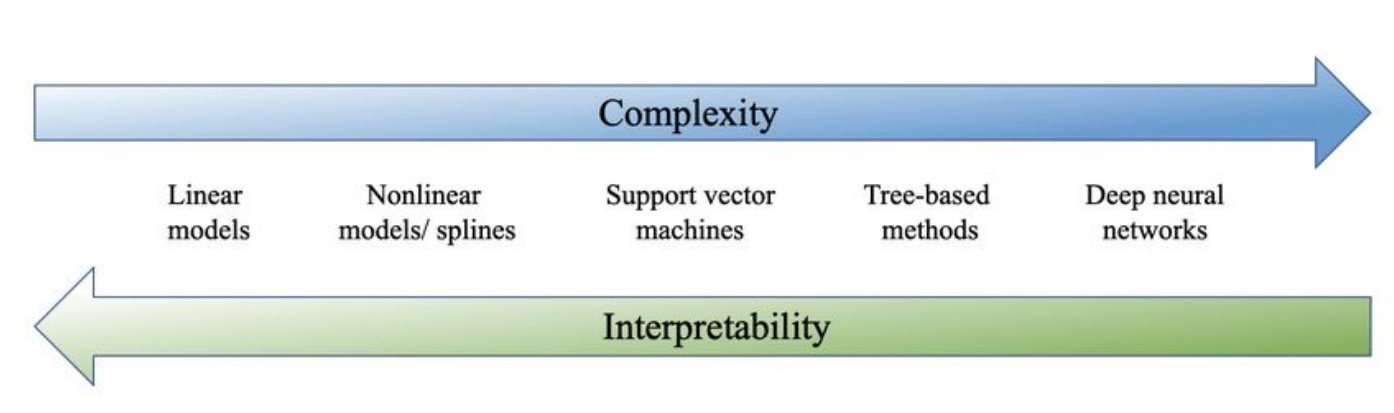
Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Roc Index
Y	Ensmbl	Ensmbl	Ensemble	class	0.975
	Reg2	Reg2	Poly Reg	class	0.941
	HPReg	HPReg	HP Regres...	class	0.931
	Neural	Neural	Neural Net...	class	0.915
	Boost	Boost	Gradient Bo...	class	0.9

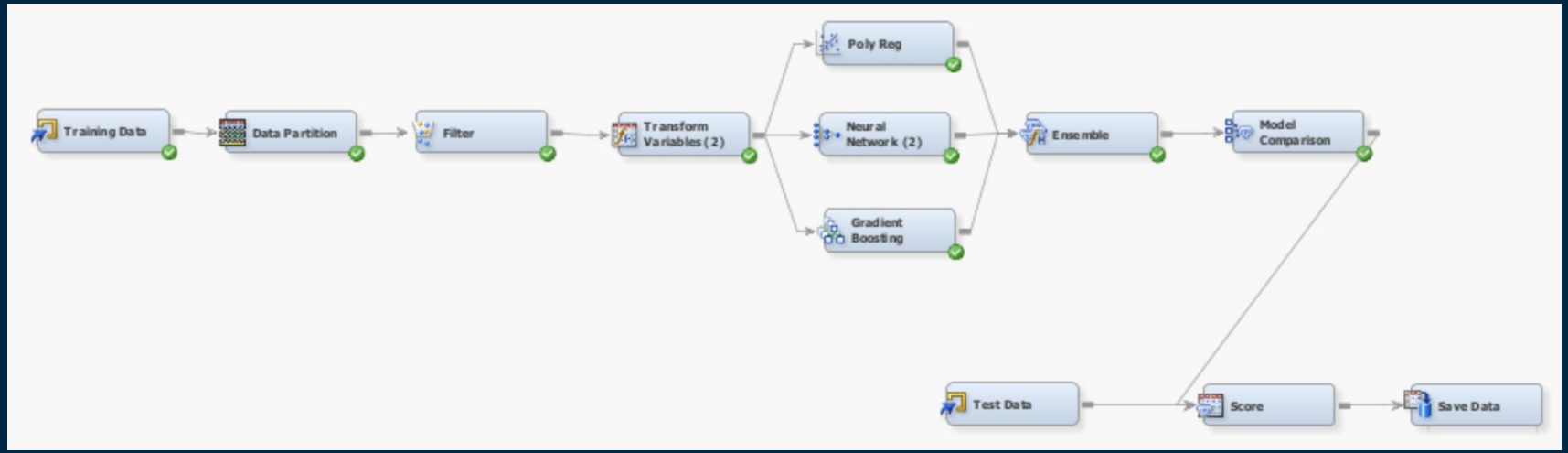
INTERPRETABILITY AND LEARNINGS

1. LG10_Attr18 (Gross profit / Total assets):
   - Coefficient: -2778.73
   - P-value: <.0001

2. LG10_Attr1*LG10_Attr19 (Net profit / Total assets and Gross profit / Sales):
   - Coefficient: -9641.38
   - P-value: <.0001



OTHER MODEL & DRAWBACK



Drawbacks: Complexity, Interpretability, hyperparameter tuning difficulty, computationally intensive, increased risk of model drift

CONCLUSION

- **SEMMA Methodology**
- **ROC across Train, Validation and Test Data**
- **Interpretability vs Complexity**
- **Hyperparameter Tuning**

THANK YOU

Why did the analyst break up with SAS Enterprise Miner?

Because every time they tried to get closer, it just kept saying "I think we need more time!"

PROPERTIES – Regressions

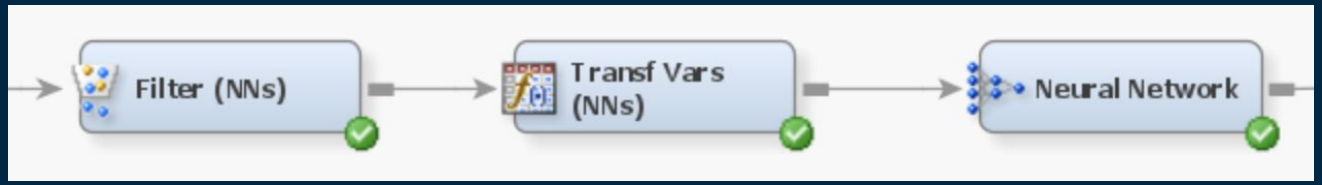
Property	Value
General	
Node ID	Filter
Imported Data	
Exported Data	
Notes	
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
Class Variables	
Class Variables	
Default Filtering Method	Rare Values (Percentage)
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percentag	0.1
Maximum Number of Levels Cu	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

Property	Value
General	
Node ID	Trans
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Formulas	
Interactions	
SAS Code	
Default Methods	
Interval Inputs	Log 10
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	8
Missing Values	Use in Search
Grouping Method	
Cutoff Value	0.2
Group Missing	No
Number of Bins	Variables
Add Minimum Value to Offset	Yes
Offset Value	1
Score	
Use Meta Transformation	Yes
Hide	Yes
Reject	Yes

Property	Value
General	
Node ID	HPReg
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	Yes
Polynomial Terms	Yes
Polynomial Degree	2
Suppress Intercept	No
Use Missing as Level	No
Modeling	
Regression Type	Logistic Regression
Link Function	Logit
Optimization Options	
Convergence Options	
Model Selection	
Selection Method	None
Selection Criterion	DEFAULT
Stop Criterion	DEFAULT
Selection Options	
Score	
Excluded Variables	Reject

Property	Value
General	
Node ID	Reg2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
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	Stepwise
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	Yes
Covariance	Yes
Correlation	Yes
Statistics	No
Suppress Output	No
Details	No
Design Matrix	No

PROPERTIES – Neural Networks

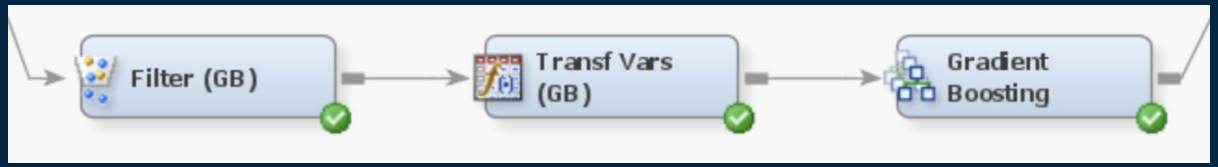


Property	Value
General	
Node ID	Filter3
Imported Data	
Exported Data	
Notes	
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
Class Variables	
Class Variables	
Default Filtering Method	Rare Values (Percentage)
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percentag	0.1
Maximum Number of Levels Cu	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	11/22/23 7:21 PM
Run ID	8ad3d091-9157-40f0-b22a-2141e06553
Last Error	
Last Status	Complete
Last Run Time	11/29/23 5:50 PM
Run Duration	0 Hr. 0 Min. 4.76 Sec.
Grid Host	
User-Added Node	No

Property	Value
General	
Node ID	Trans2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Formulas	
Interactions	
SAS Code	
Default Methods	
Interval Inputs	Log 10
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	8
Missing Values	Use in Search
Grouping Method	
Cutoff Value	0.2
Group Missing	No
Number of Bins	Variables
Add Minimum Value to Offset	Yes
Offset Value	1
Score	
Use Meta Transformation	Yes
Hide	Yes
Reject	Yes

Property	Value
General	
Node ID	Neural
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	12345
Model Selection Criterion	Profit/Loss
Suppress Output	No
Score	
Hidden Units	Yes
Residuals	Yes
Standardization	No
Status	
Create Time	11/25/23 9:11 AM
Run ID	886e6824-11a2-461f-8756-1c11342509
Last Error	
Last Status	Complete
Last Run Time	11/29/23 6:04 PM
Run Duration	0 Hr. 0 Min. 15.48 Sec.
Grid Host	
User-Added Node	No

PROPERTIES – Gradient Boosting



General	
Node ID	Filter2
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	Rare Values (Percentage
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percentag	0.1
Maximum Number of Levels Cu	25
Interval Variables	
Interval Variables	
Default Filtering Method	None
Keep Missing Values	Yes
Tuning Parameters	
Score	
Create Score Code	Yes
Update Measurement Level	No

Property	Value
General	
Node ID	Trans3
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Formulas	
Interactions	
SAS Code	
Default Methods	
Interval Inputs	None
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	8
Missing Values	Use in Search
Grouping Method	
Cutoff Value	0.2
Group Missing	No
Number of Bins	Variables
Add Minimum Value to Offset	Yes
Offset Value	1
Score	
Use Meta Transformation	Yes
Hide	Yes
Reject	Yes

Property	Value
General	
Node ID	Boost
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Series Options	
N Iterations	150
Seed	12345
Shrinkage	0.05
Train Proportion	70
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	7000
Node Sample	20000
Subtree	
Assessment Measure	Decision
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes