Introduction to Machine Learning

Melodie Rush

SAS Customer Success – Principal Data Scientist

Connect with me:

LinkedIn: https://www.linkedin.com/in/melodierush

Twitter: @Melodie_Rush

Agenda

- What is Machine Learning?
- Machine Learning Terminology
- Intro to ML Modeling Algorithms
- Machine Learning in SAS Viya

Machine Learning

HEAVILY HYPED SELF-DRIVING CAR
ONLINE RECOMMENDATION OFFERS
CHAT BOTS
FRAUD DETECTION
TARGETED ADS

What is Machine Learning?

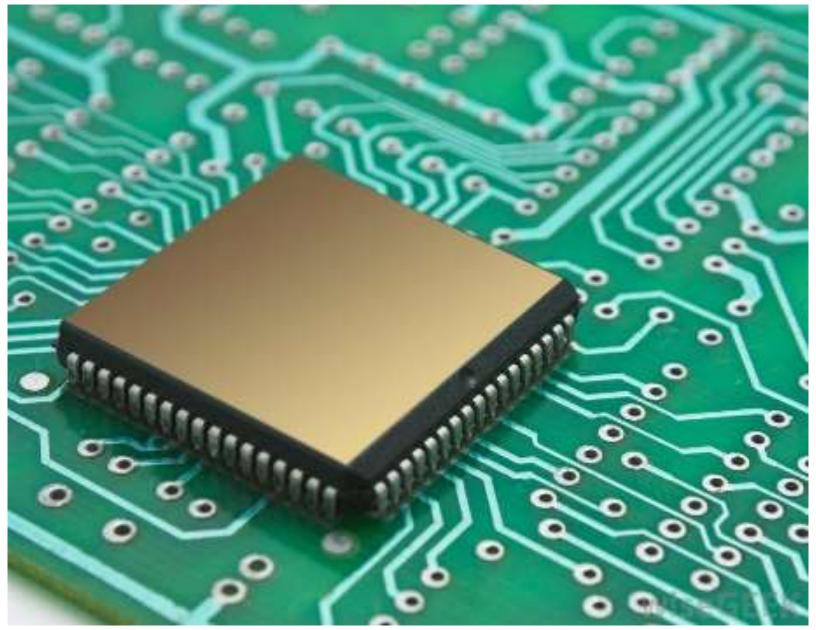
Definition

- Automatic
- Adaptive

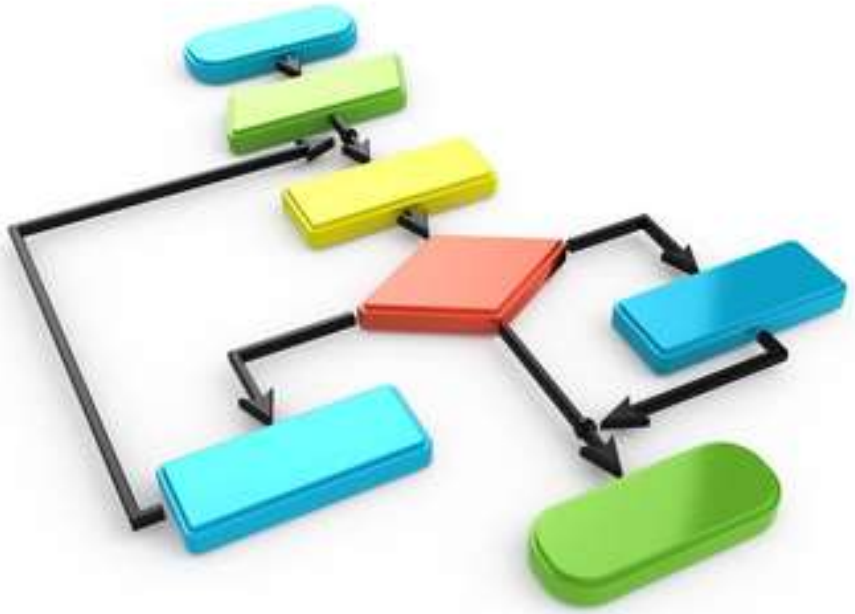
Using iterative processes, machine learning builds models that ***automatically adapt*** *with little or no human intervention.*

Why is it so important now?

Data



Computing Power

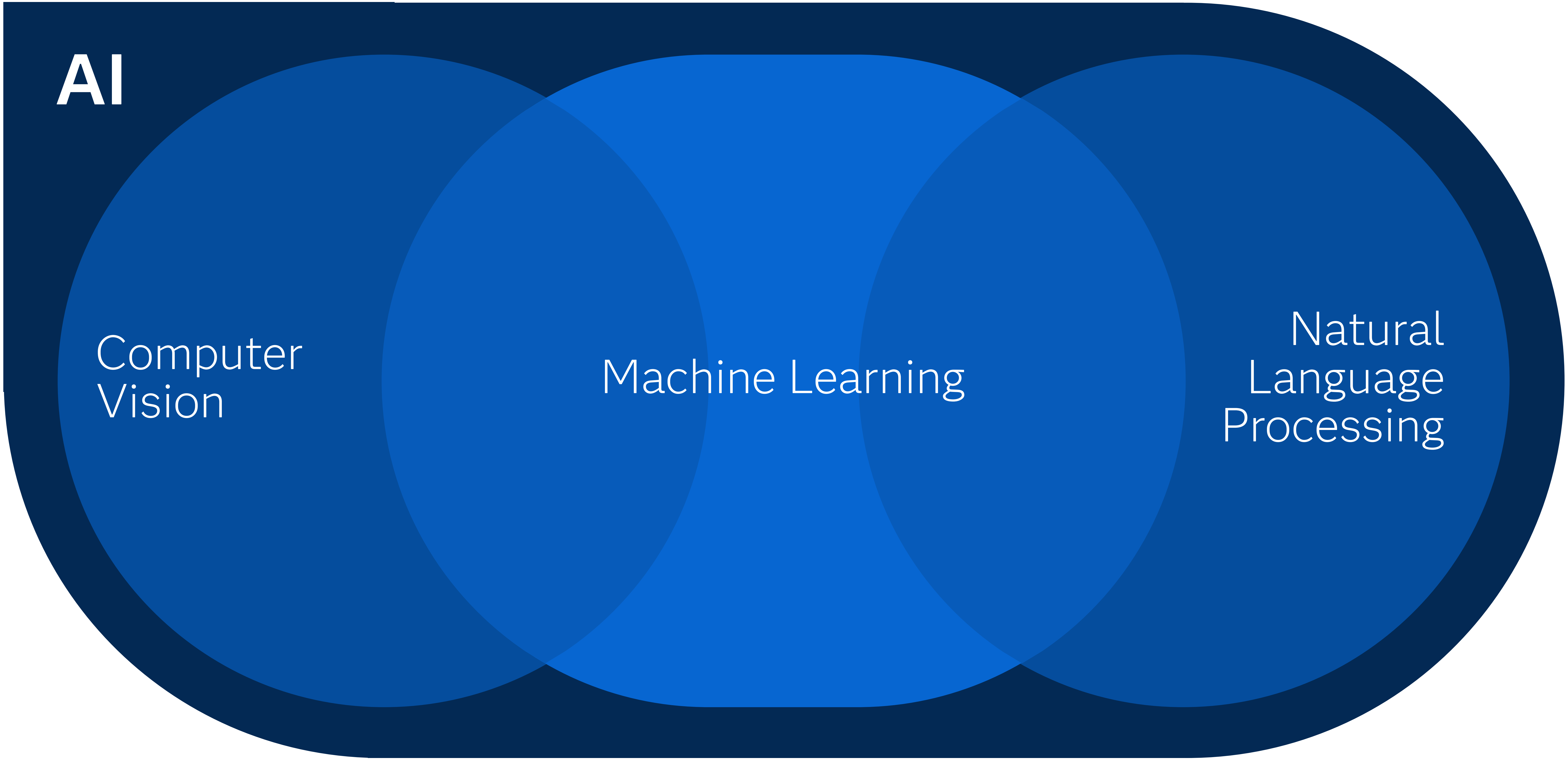


Algorithms

What is Artificial Intelligence?

AI

Science of designing computer systems to support and accelerate human decisions and actions.


AI
Computer Vision
Machine Learning
Natural Language Processing

What is Machine Learning?

ML is a branch of artificial intelligence that **automates** the building of systems that learn from data, identify patterns, and predict future results – with **minimal human intervention**.

AI or ML?

What’s the Difference

AI systems perform tasks that typically require human-level intelligence

- Understanding Language
- Recognizing images and patterns
- Making Decisions
- Learning from the past

Machine Learning uses data & algorithms to learn and make decisions

- ML may be part of the brains in an AI system, or it may be used in a stand-alone usage
- Generally, we think of predictive modeling

Terminology

In Machine Learning

Terminology

Machine learning terms versus inferential statistics terms

What are all these archaic, outmoded and confusing terms?

What are all these new fangled and confusing terms?

- Feature
- Input
- Target
- Object

- Variable
- Independent Variable
- Dependent Variable
- Observation

Terminology

What are Machine Learning terminology?

- In statistics we predict a Y or a dependent variable.
- In data mining, Y is called a target.
- In machine learning, a target is called a label.
- In statistics and data mining our inputs are called X’s.
- In machine learning our inputs are called features.
- In statistics and data mining we transform our X’s.
- In machine learning we do feature creation.

AI or ML?

What’s the Difference

AI systems perform tasks that typically require human-level intelligence

- Understanding Language
- Recognizing images and patterns
- Making Decisions
- Learning from the past

Machine Learning uses data & algorithms to learn and make decisions

- ML may be part of the brains in an AI system, or it may be used in a stand-alone usage
- Generally, we think of predictive modeling

How do Models Learn?

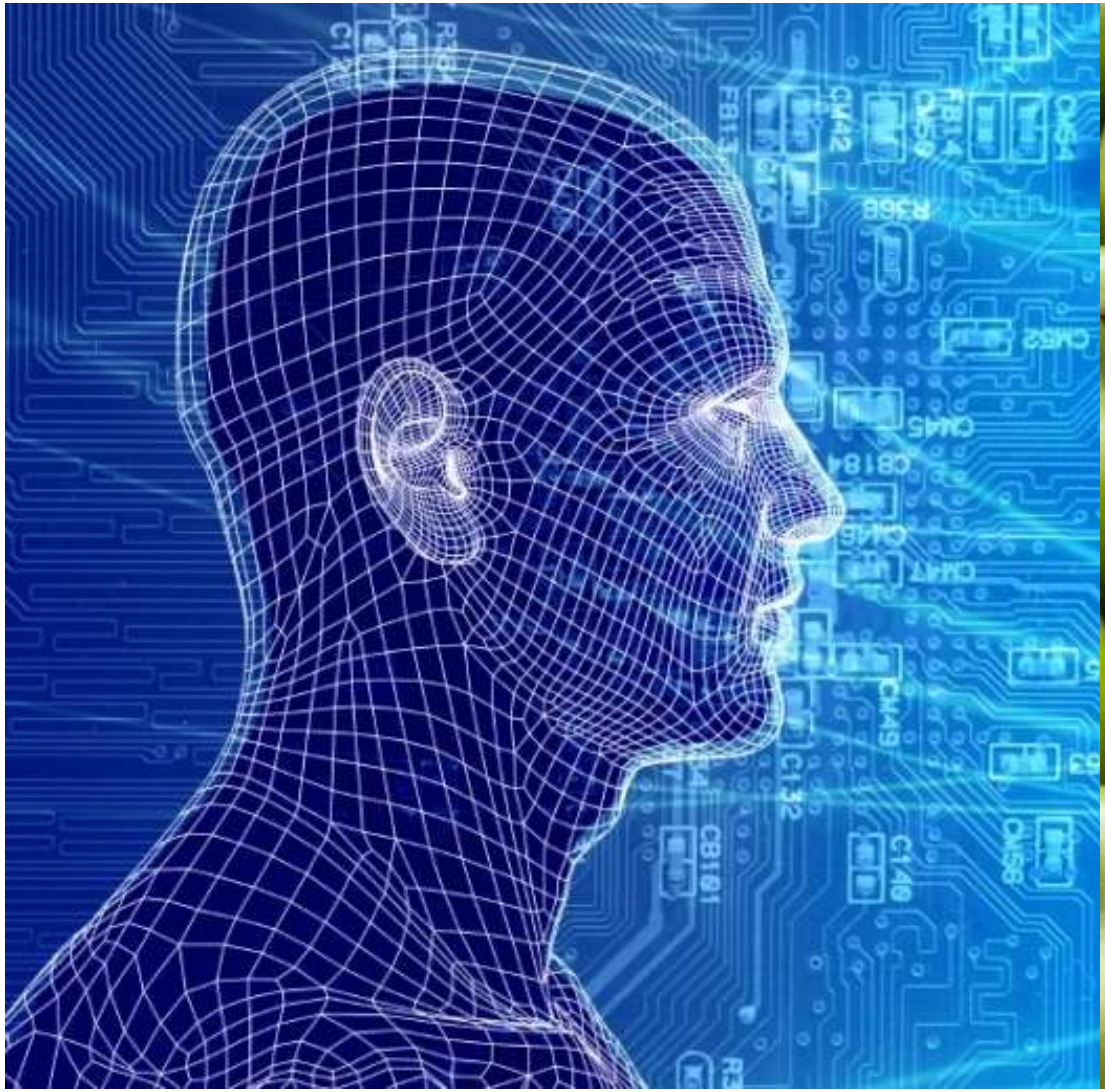
Distinguish apple from orange

How do Models Learn?

Distinguish Granny Smith apple from Fuji apple

How do Models Learn?

Finding the rotten apple



How do Models Learn – New Data

Predictions

How do Models Learn – New Data

Predictions

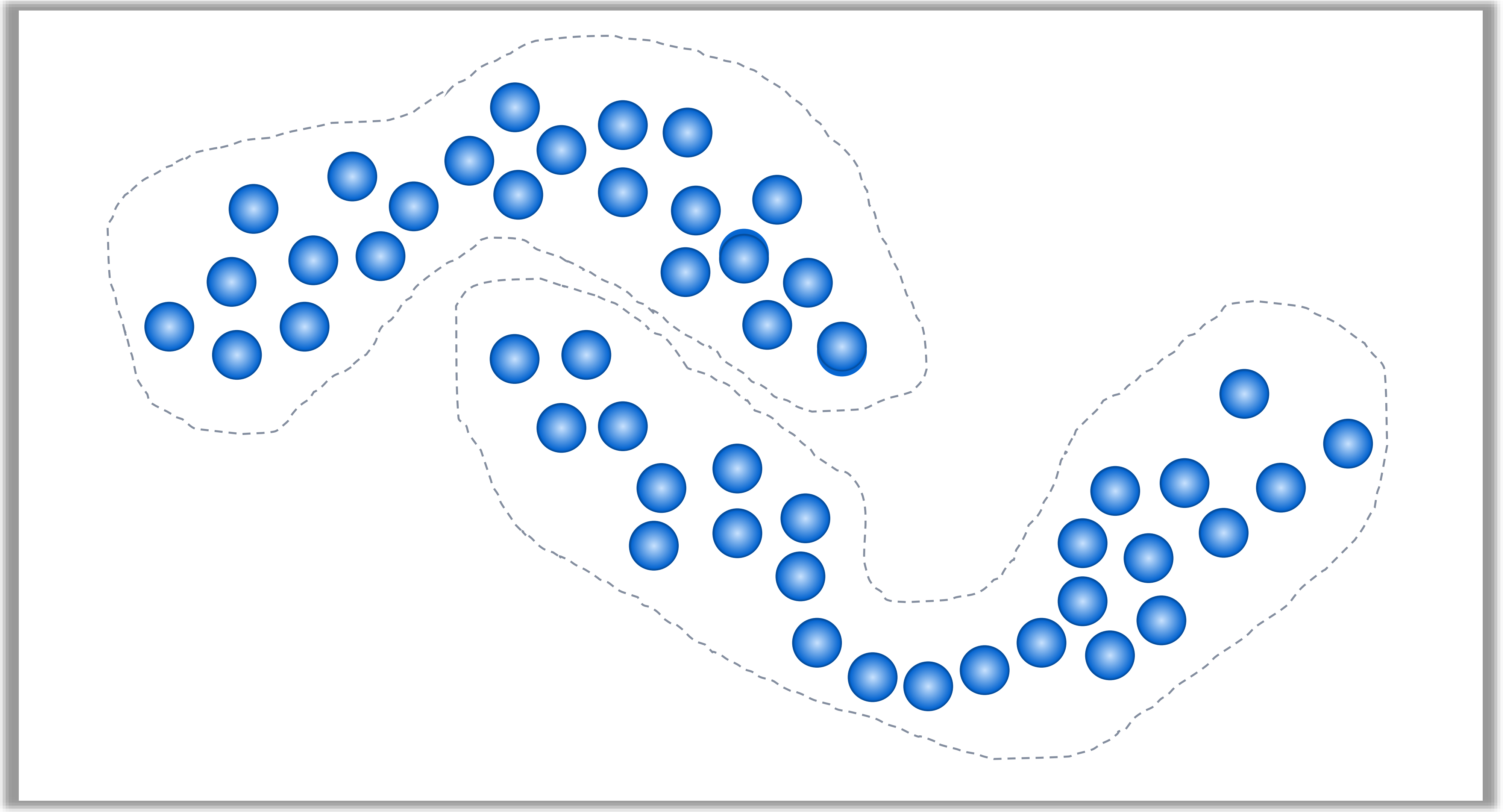
How do Models Learn – New Data

Predictions

How Does Machine Learning Work?

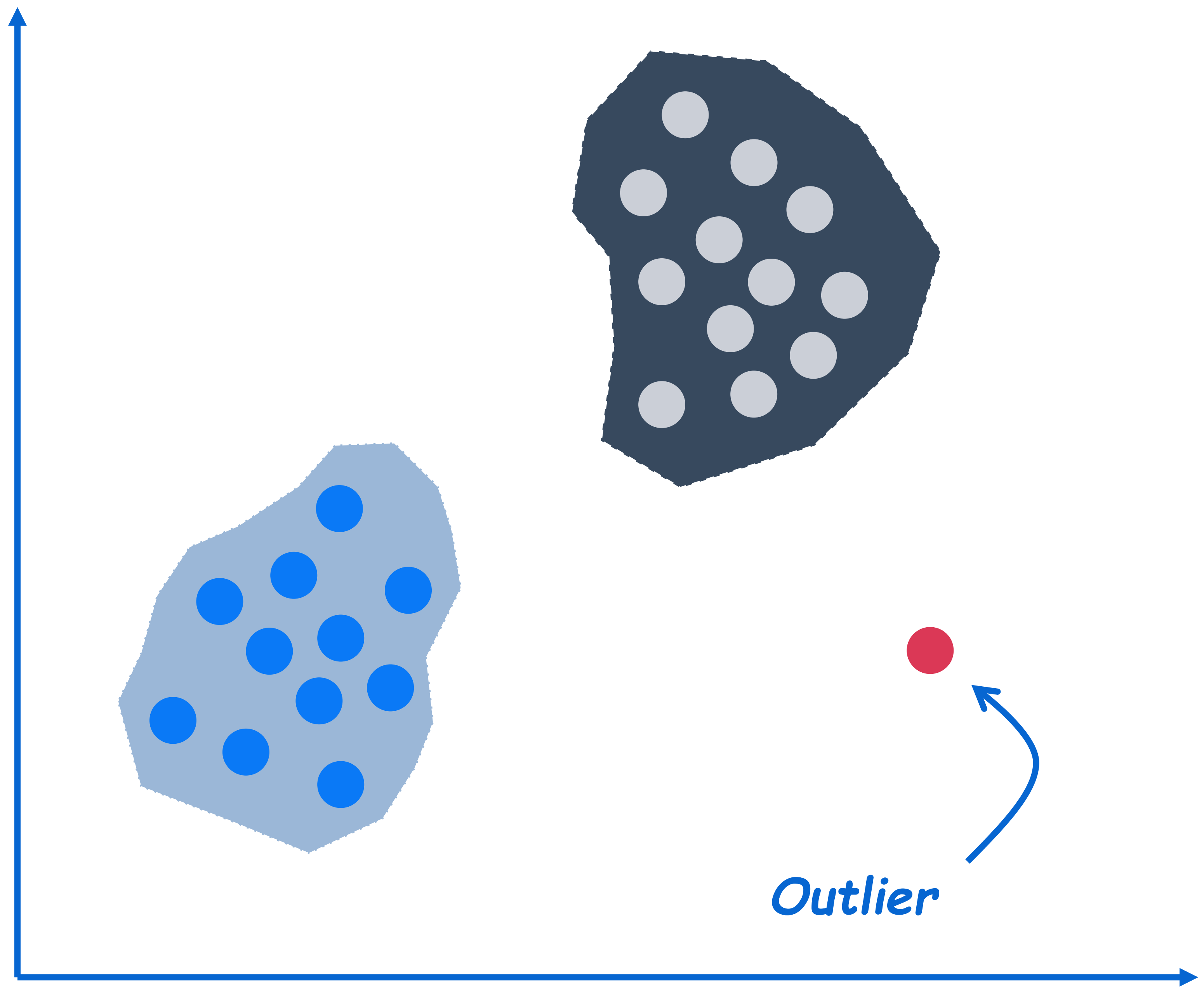
Unsupervised Learning

Trained on unlabeled examples



Machine Learning

Unsupervised Learning



Clustering

Groups of similar data
(e.g. related products on
supermarket basket)

Anomaly Detection

Identifying outliers
(e.g. Abnormal credit card
transactions)

Machine Learning

Unsupervised Learning Use Cases

Segmentation

Recommendation Engines

Anomaly Detection

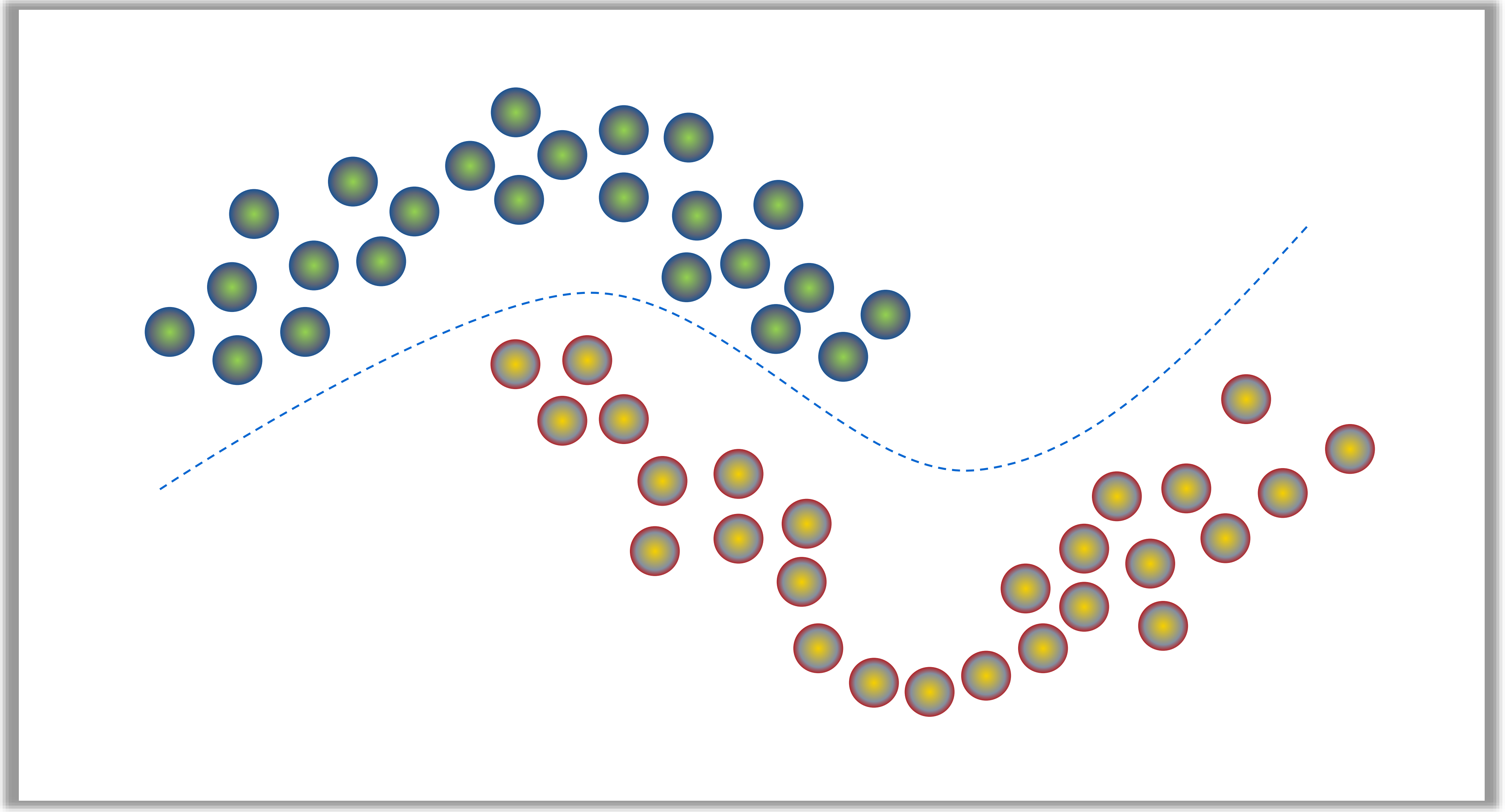
Predictive Maintenance

Dimensionality Reduction

How Does Machine Learning Work?

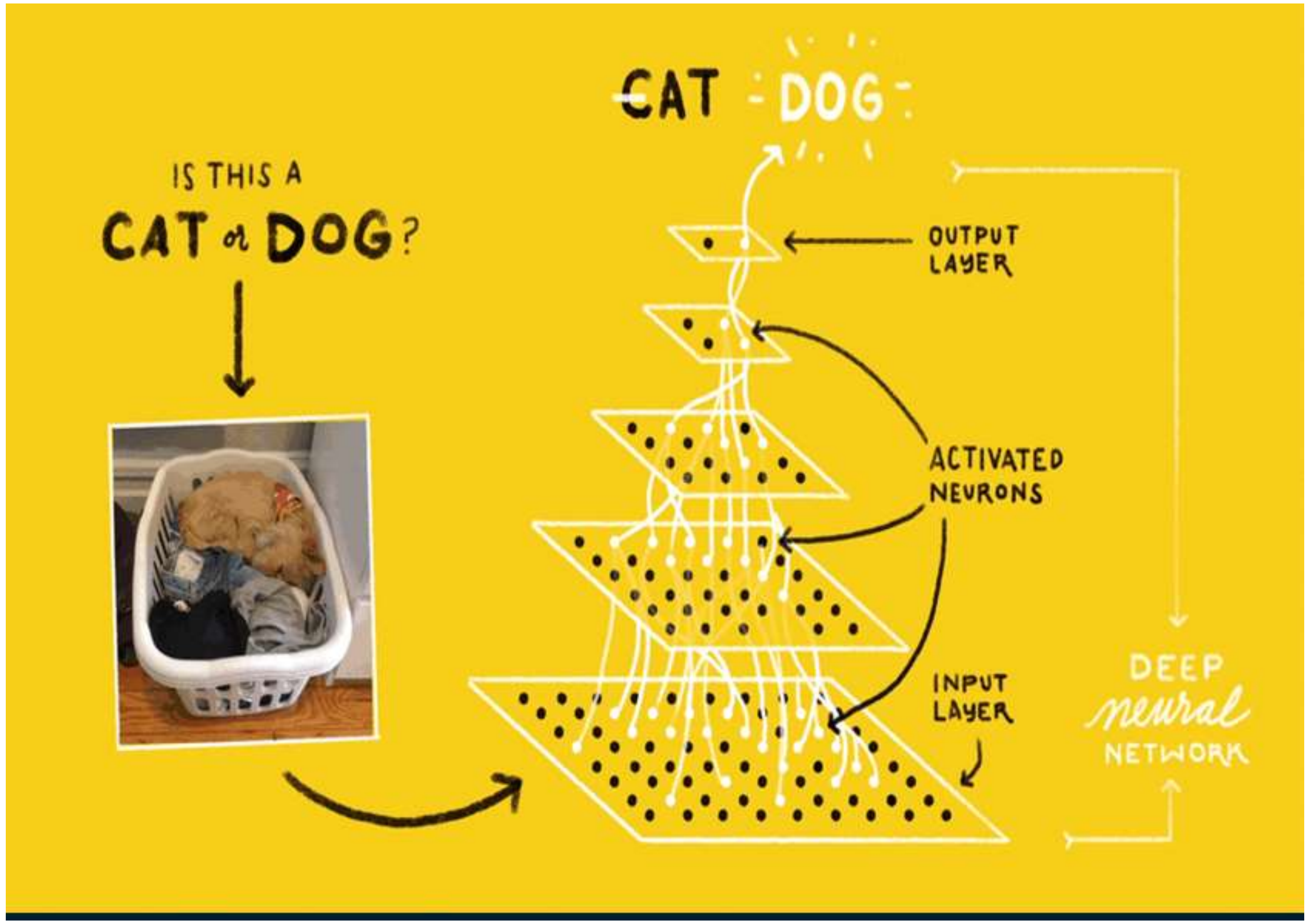
Supervised Learning

Trained on labeled examples



Machine Learning

Supervised Learning



Regression

Predict a numerical value
(e.g. price of a house,
demand for milk)

Classification

Predict a label or future
event
(e.g. Cat or Dog, Probability
of loan default)

Machine Learning

Supervised Learning Use Cases

Risk Modeling

Fraud Detection

Forecasting

Customer Retention

Document Classification

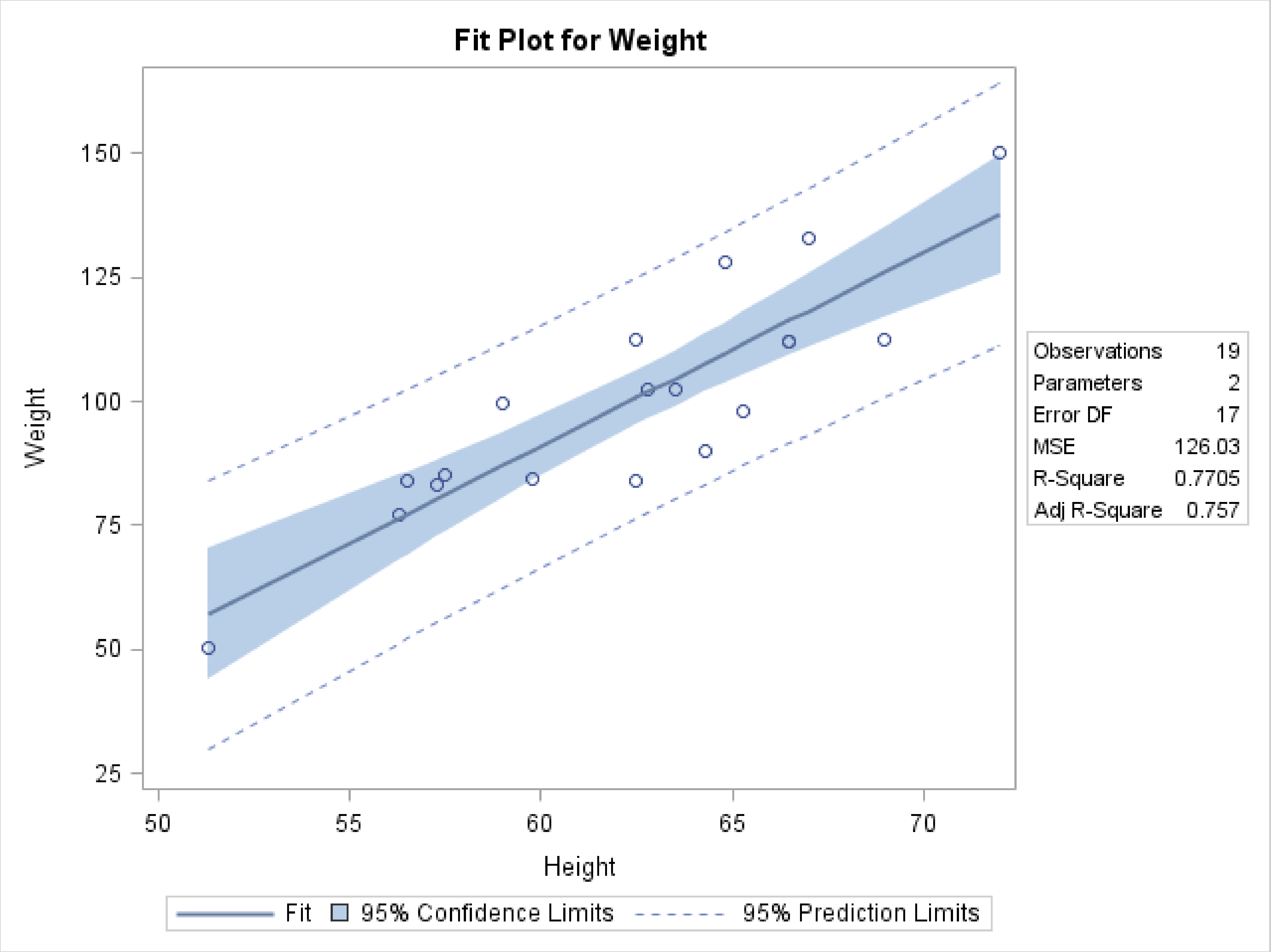
Machine Learning Algorithms

Available in SAS Viya

Regression

What Is It?

- Used to identify the relationship between a dependent variable and one or more independent variables
- Many types – linear, logistic, quantile, polynomial, stepwise, ridge, lasso, ElasticNet, etc…
- Oldie but goodie



Observations	19
Parameters	2
Error DF	17
MSE	126.03
R-Square	0.7705
Adj R-Square	0.757

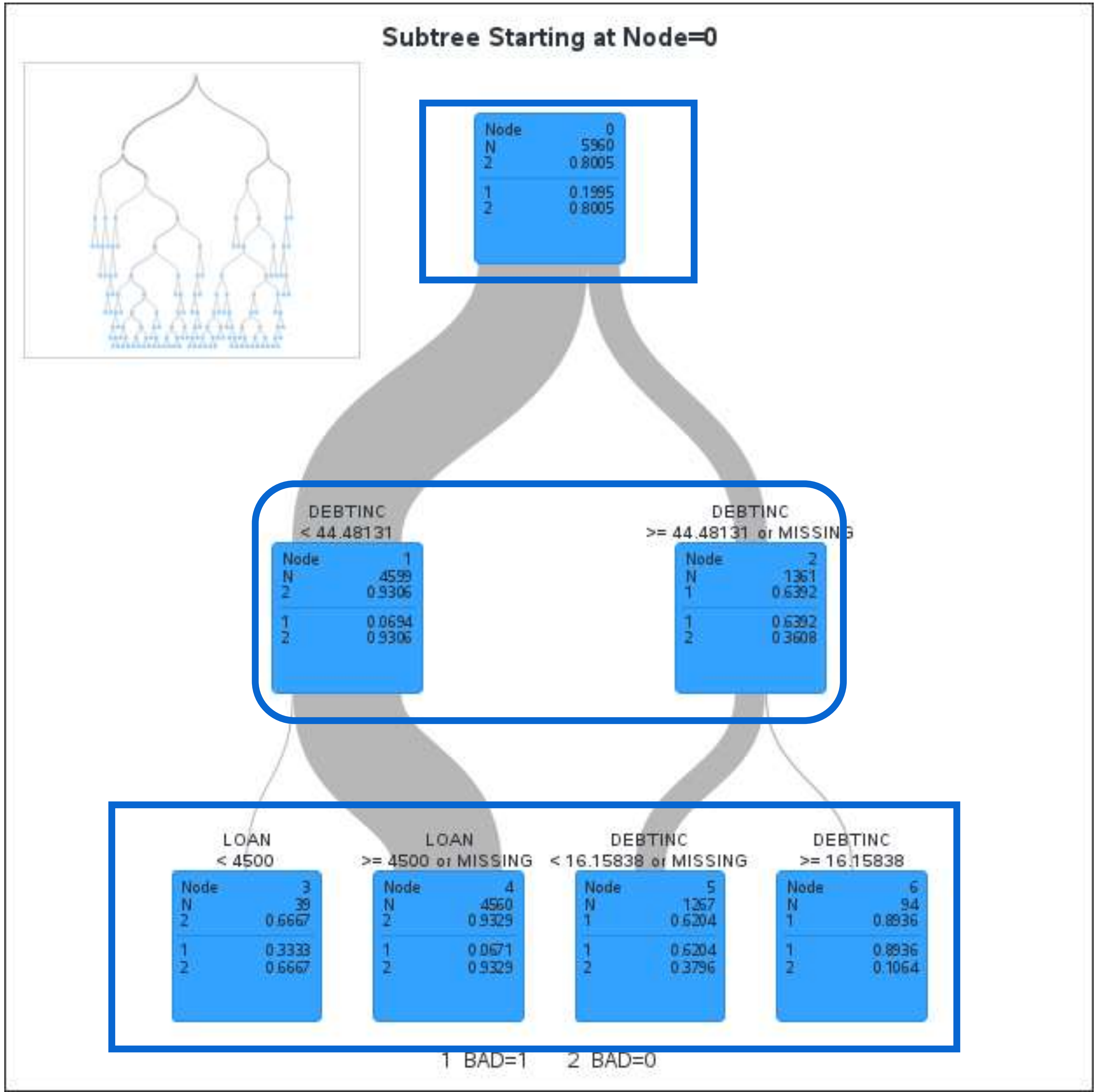
Decision Trees

What Is It?

- Linear separation of data using “if then else” logic
- Separation is performed via an exhaustive search of splitting points for each variable.
- Many different architectural variations based on the above architecture
- Users might refer to them as
  - CHAID Trees
  - CART Trees
  - C4.5 Trees
  - C5.0 Trees.
  - Each of the above is simply a variation on the tree a

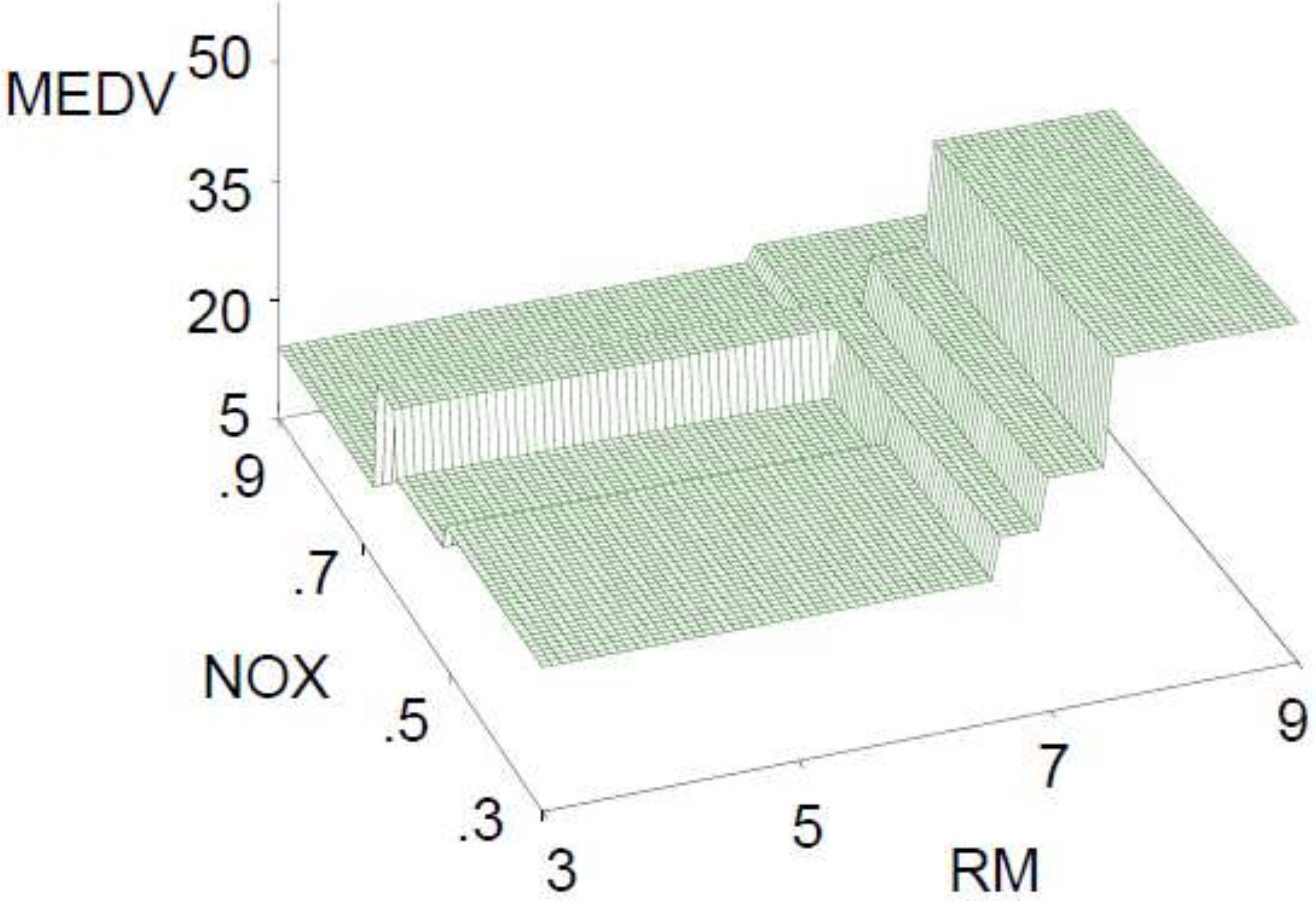
Decision Tree

- Easy to Visualize



Decision Trees

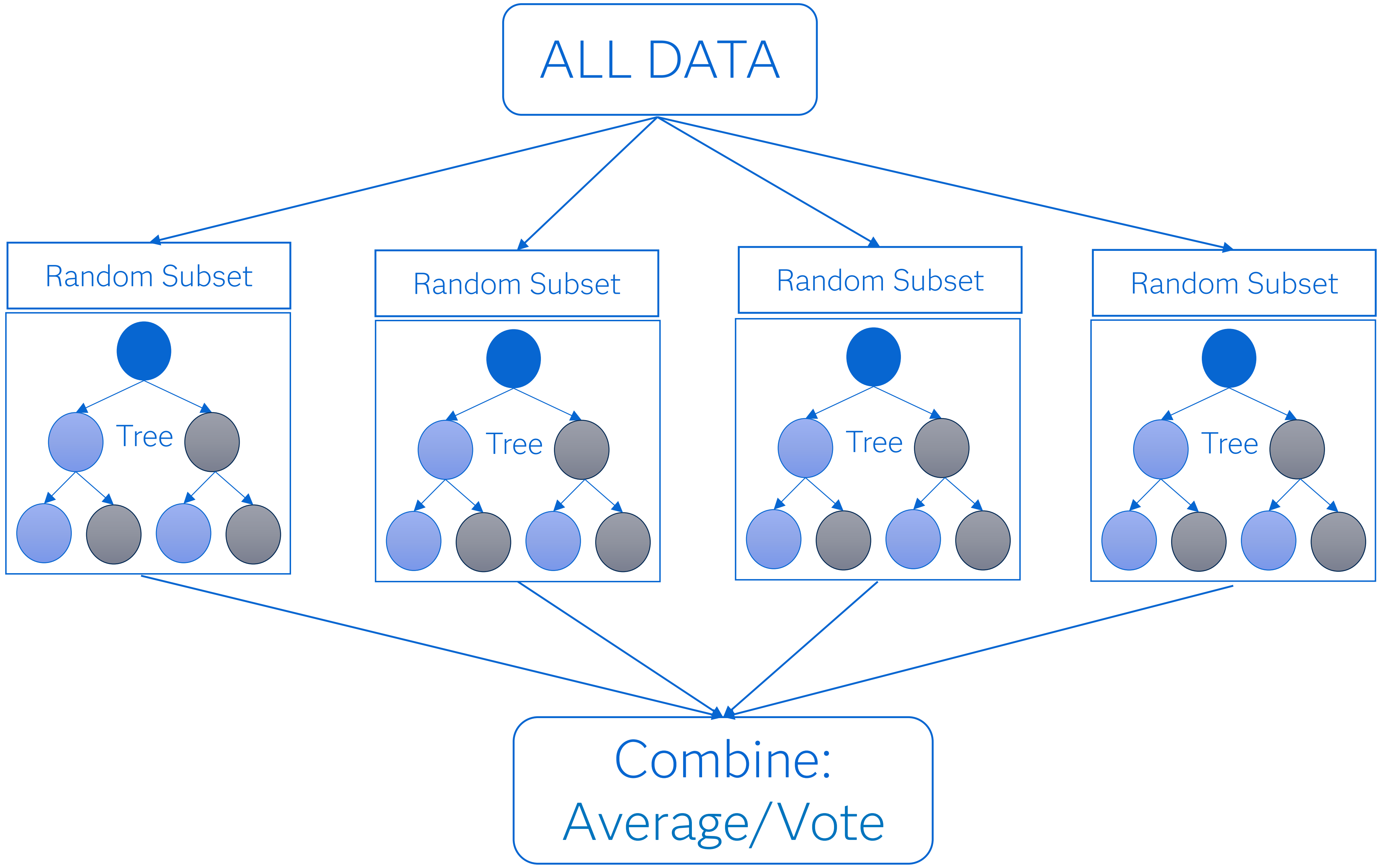
Multivariate Step Function



Random Forest

What Is It?

- A combination of several "decision trees."
- A random forest consists of a forest of fully trained decision trees.
- The random forest averages the output of all the decision trees in the "forest."



Random Forest

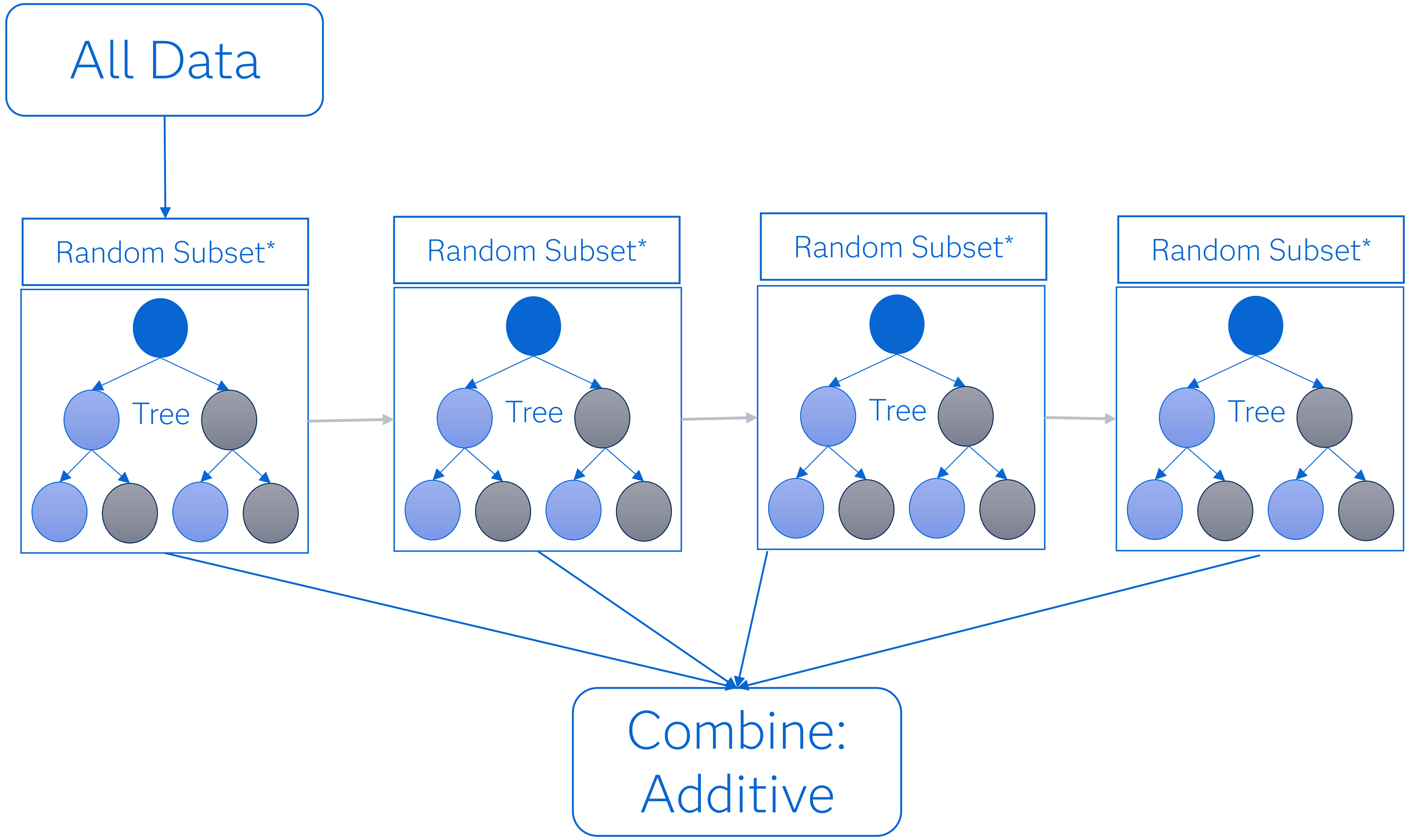
Algorithm

- Select a number of trees in the random forest.
- For each tree in the forest, use the following split algorithm:
    - Select a random sample of data.
    - Select a random subset of variables.
    - Determine the best split from the sample of data and the sample of variables.
    - Keep selecting random data and random subsets of variables until the maximum number of trees is trained.
- When all the trees are built, the prediction is the average of all trees.

Gradient Boosting

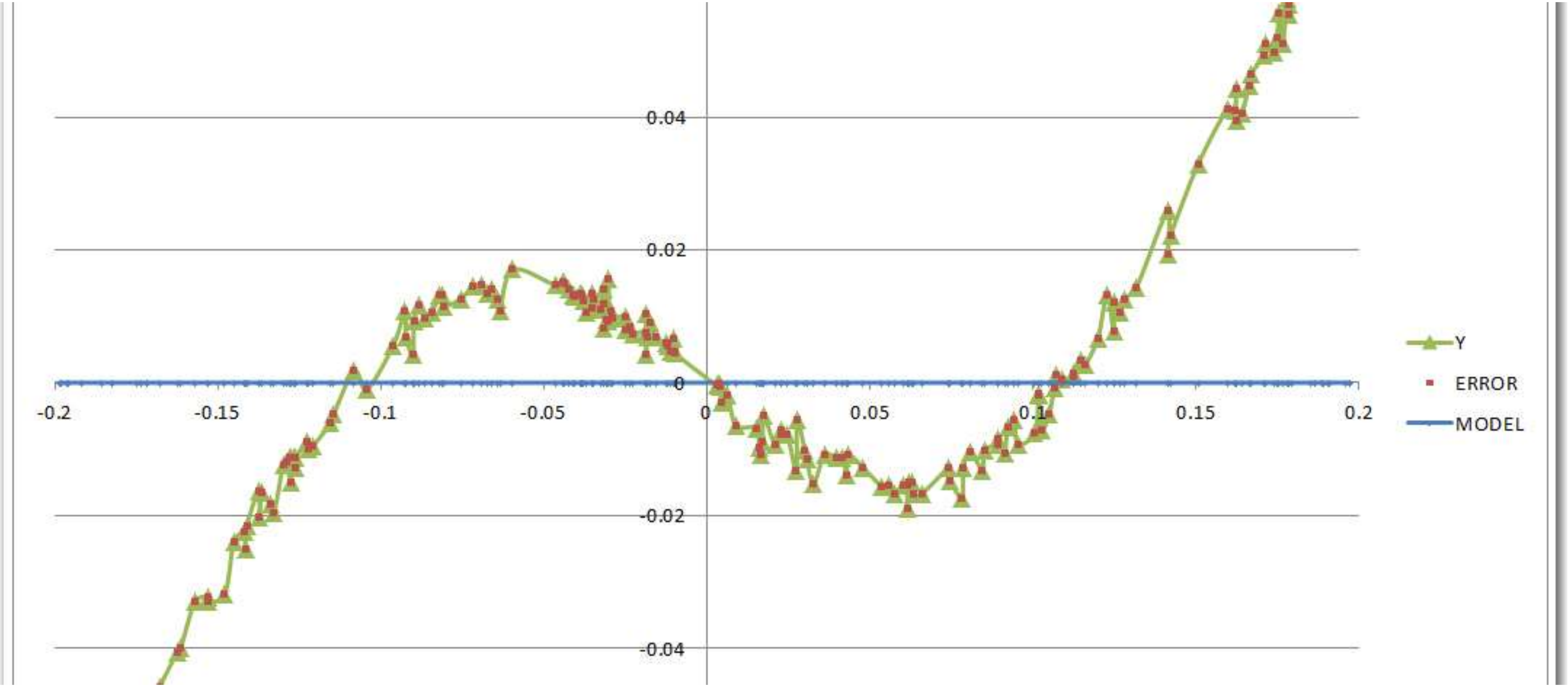
What Is It?

- A combination of several "decision trees."
- Gradient boosting consists of a **forest** of **small** decision trees ("**shrubs**", "stumps").
- Each **shrub** is poor at predicting target, but each subsequent shrub tries to fit the remaining error.
- Eventually converges to good solution.



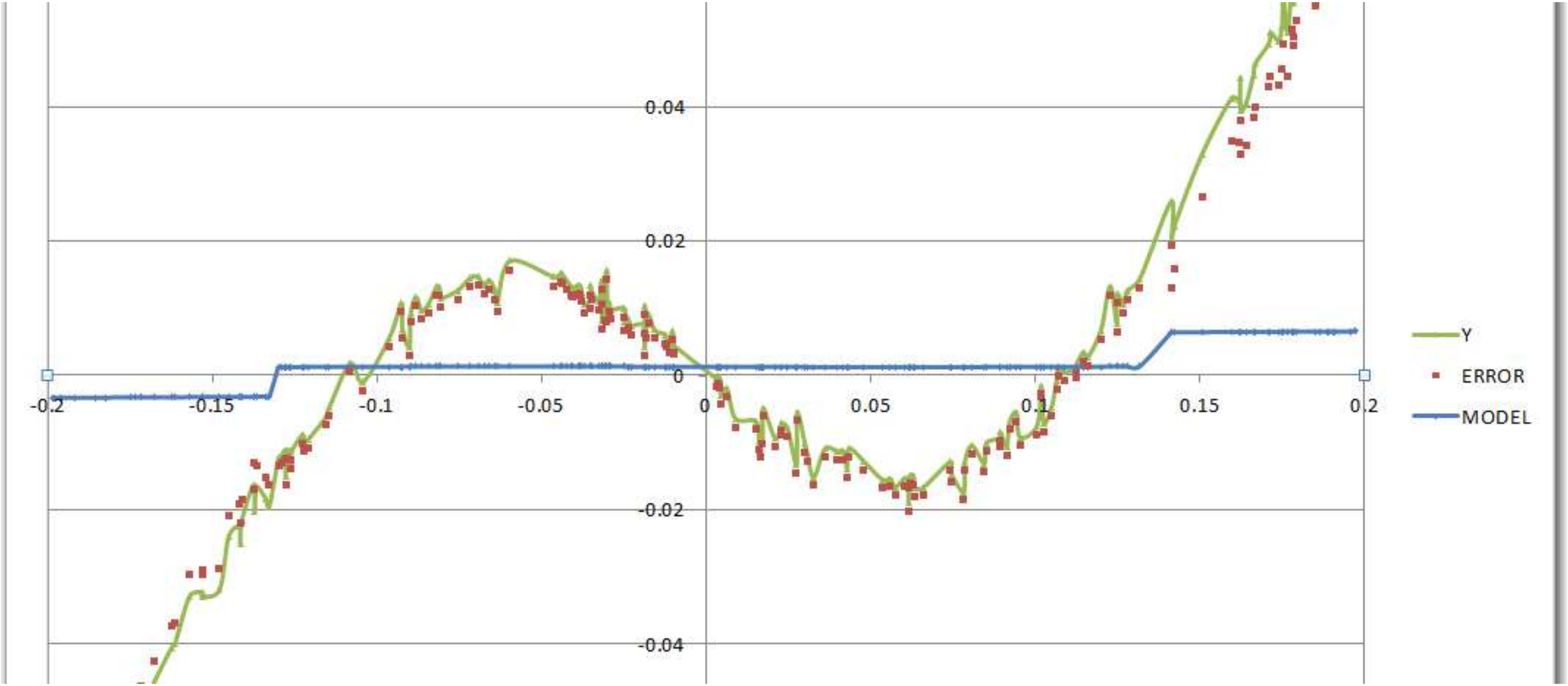
Gradient Boosting

Example: Iterations=0



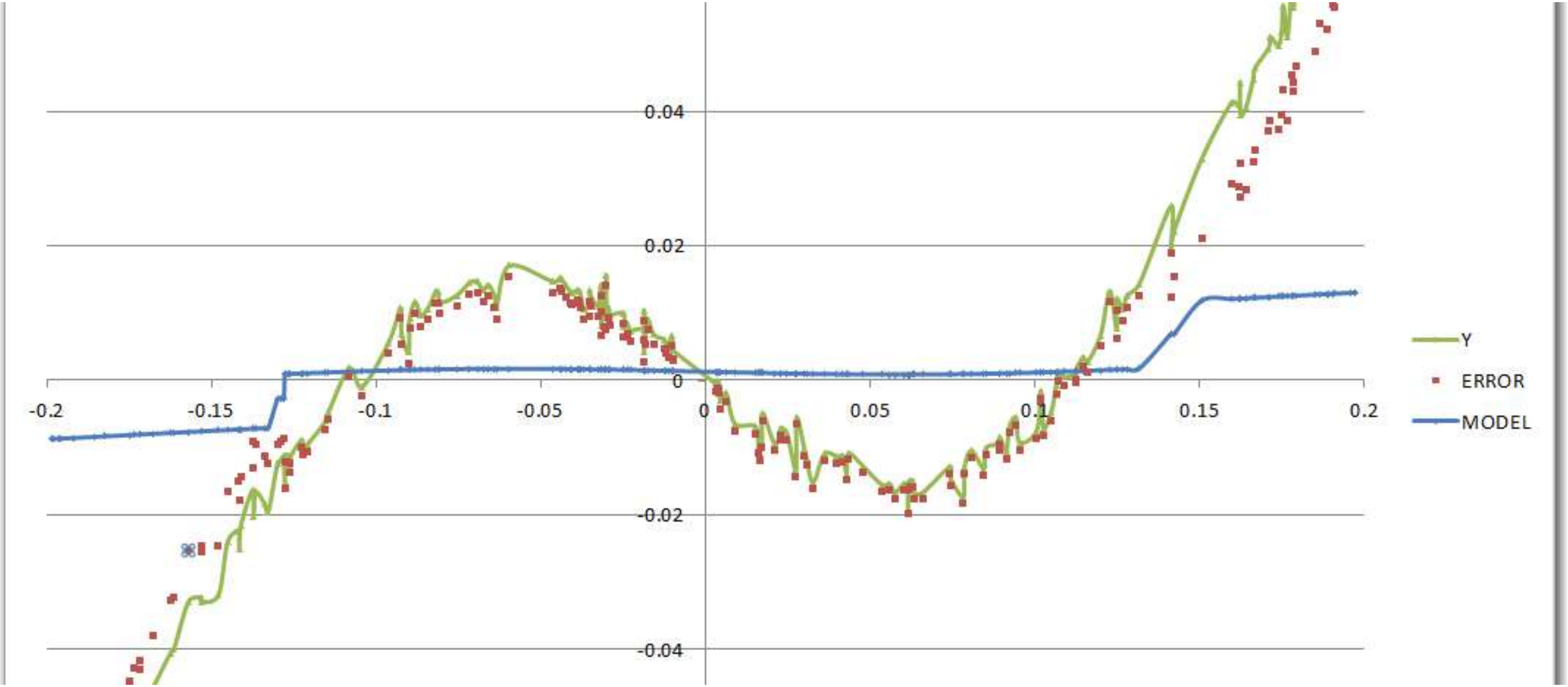
Gradient Boosting

Example: Iterations=1



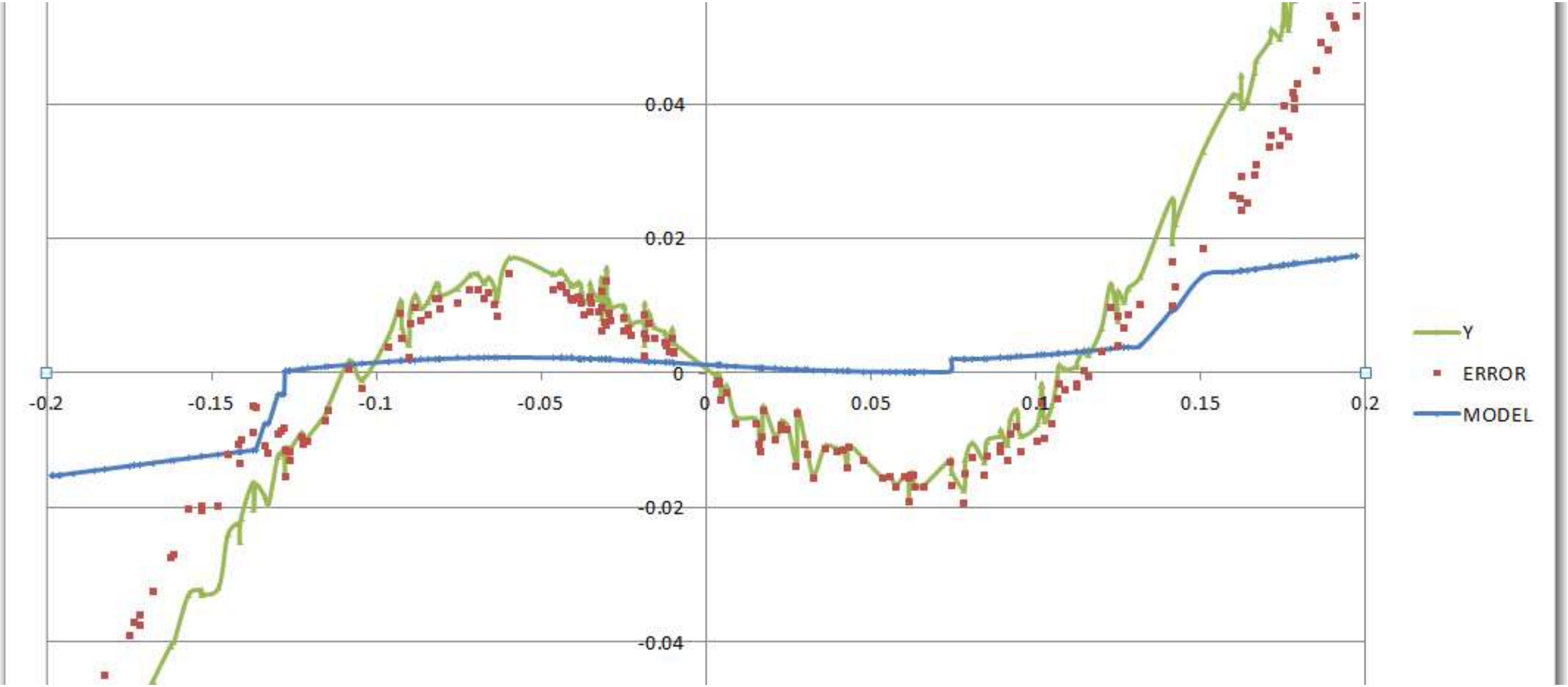
Gradient Boosting

Example: Iterations=10



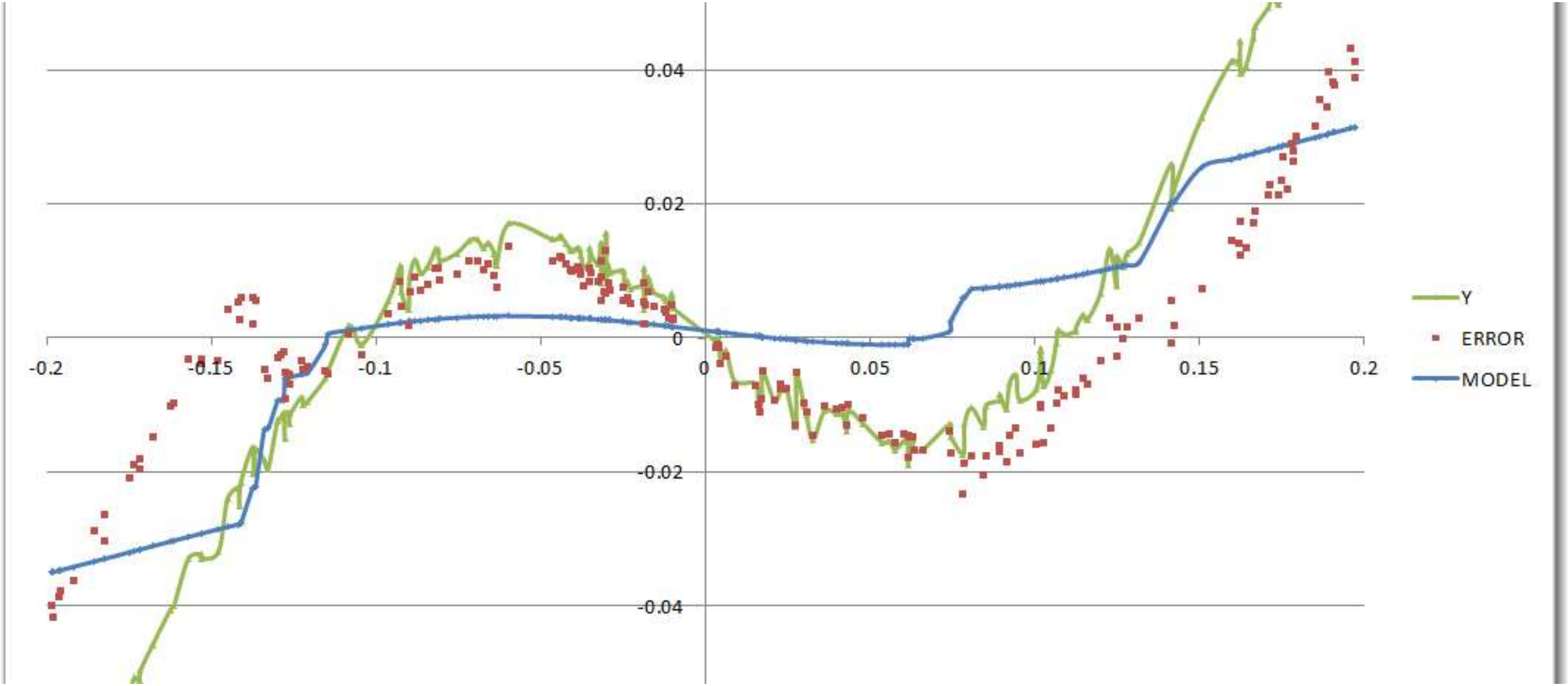
Gradient Boosting

Example: Iterations=25



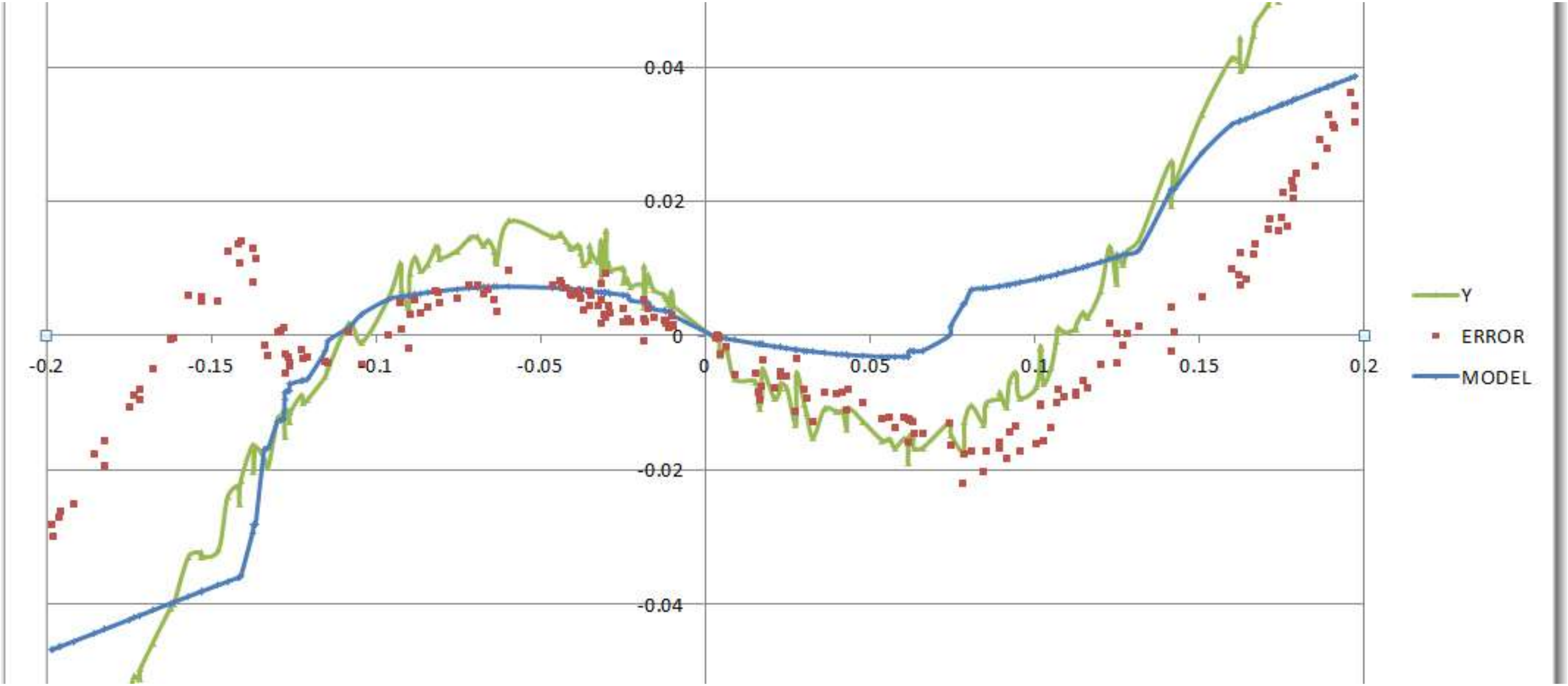
Gradient Boosting

Example: Iterations=50



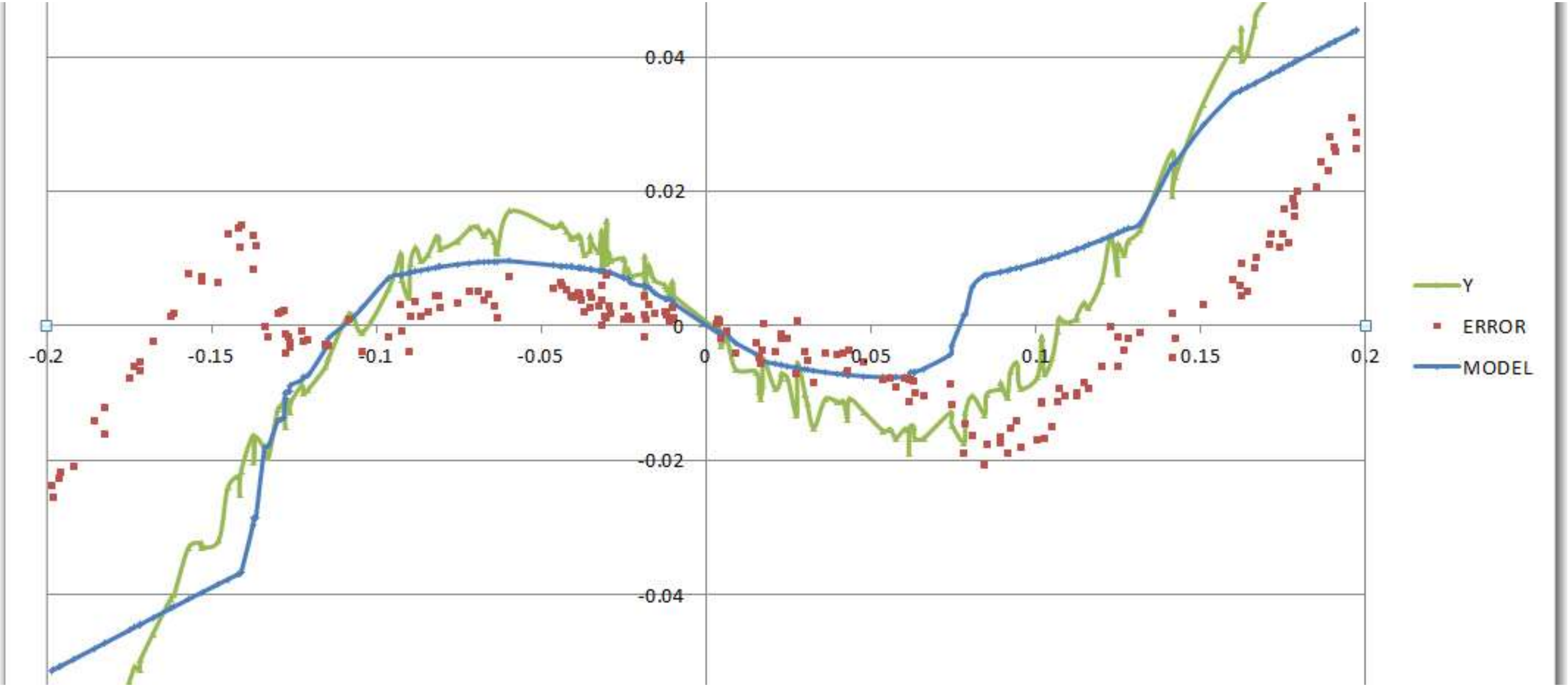
Gradient Boosting

Example: Iterations=75



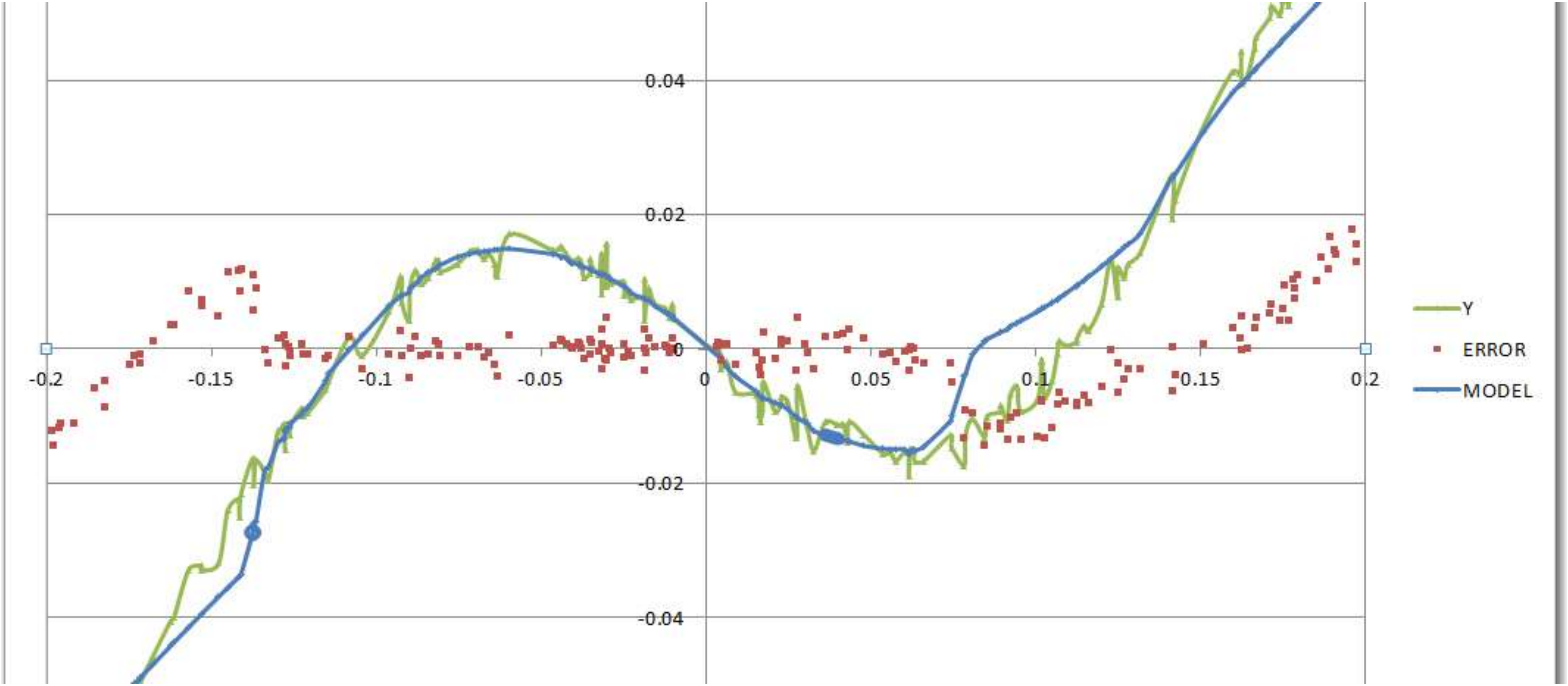
Gradient Boosting

Example: Iterations=100



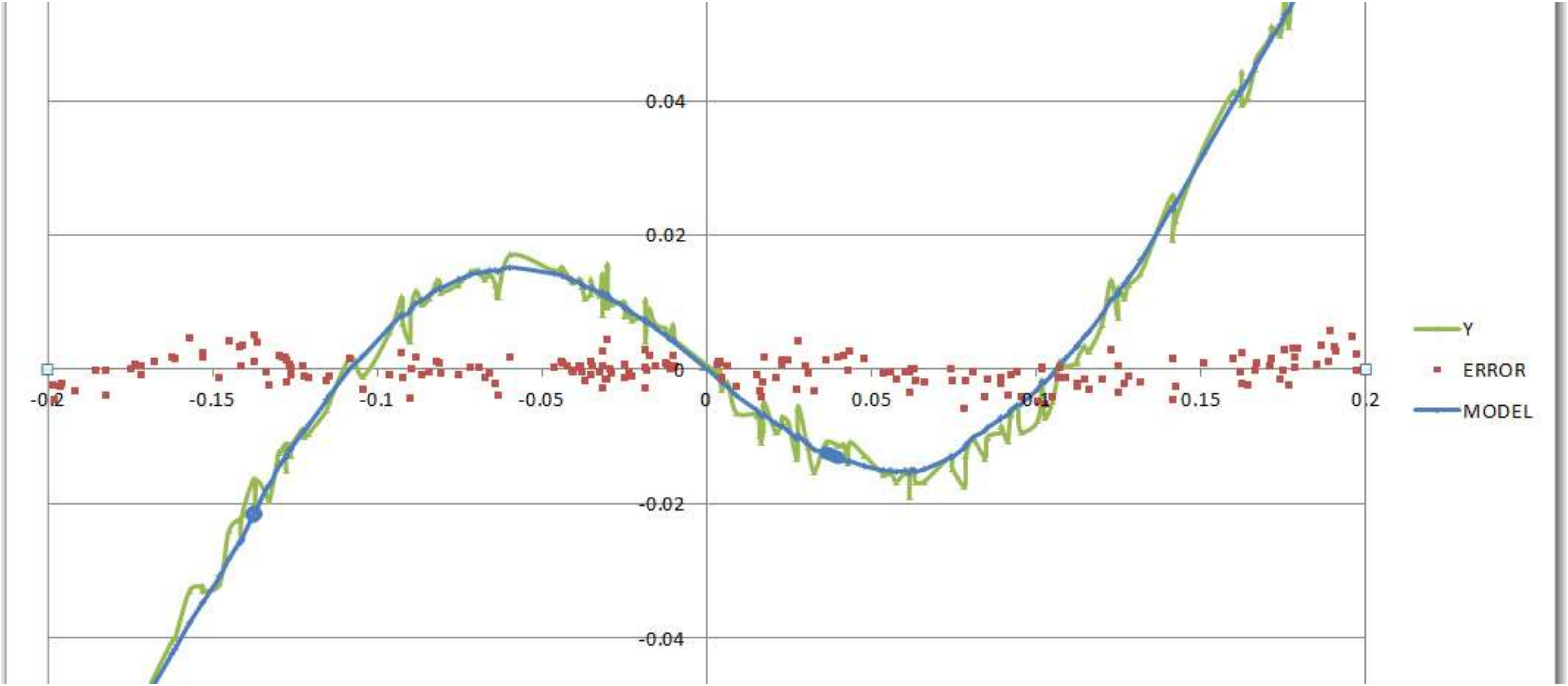
Gradient Boosting

Example: Iterations=200

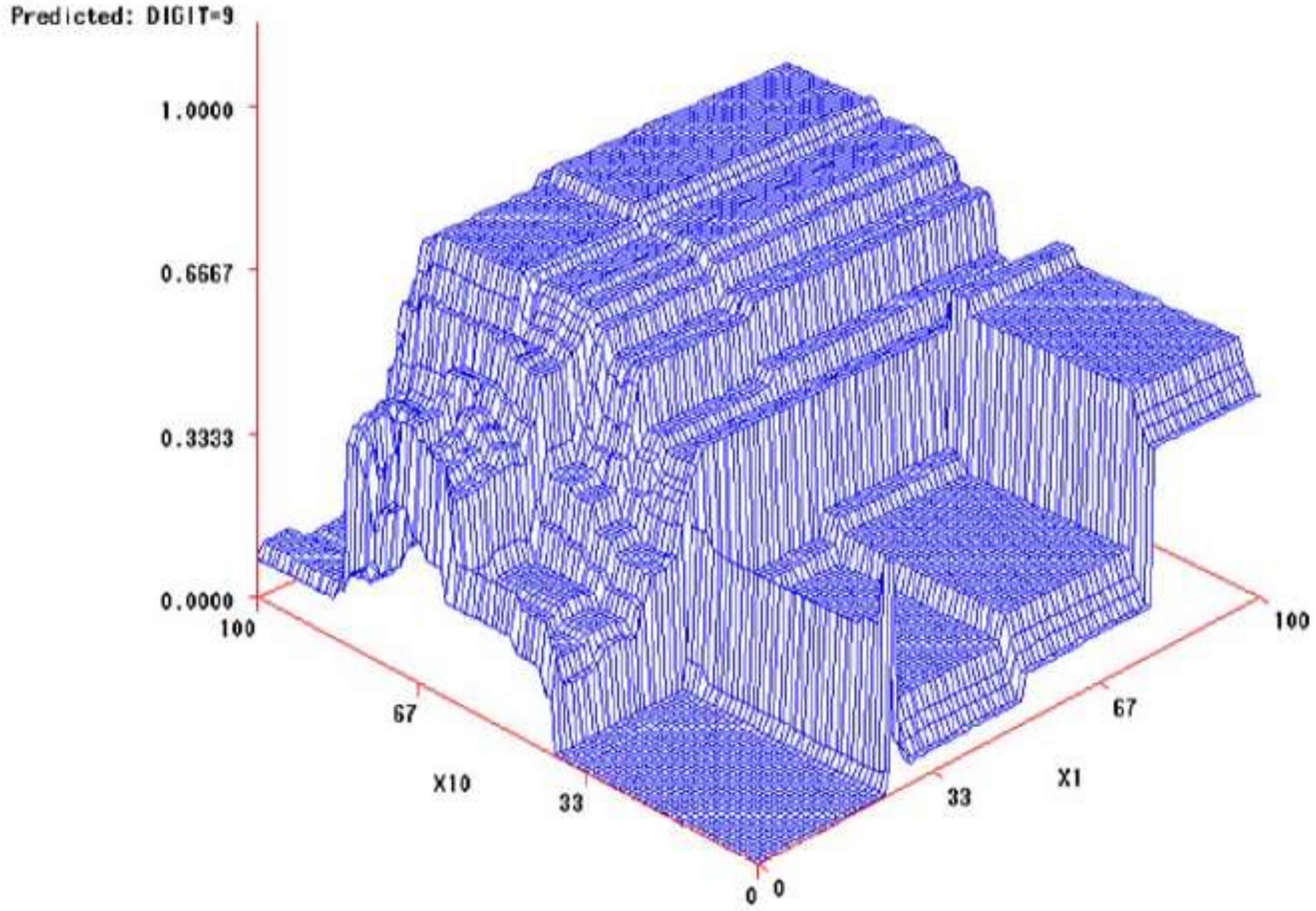


Gradient Boosting

Example: Iterations=300



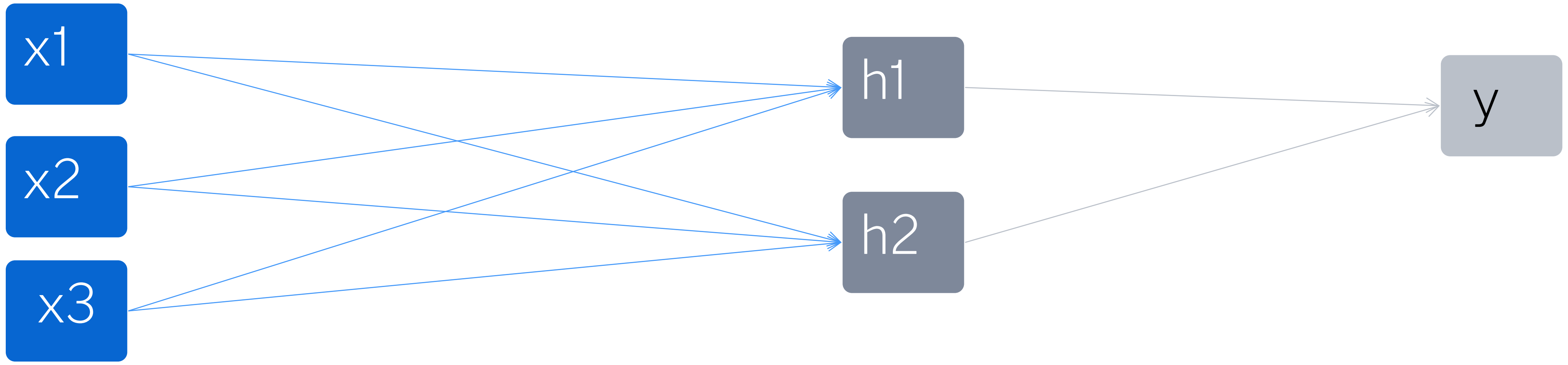
Gradient Boosting



Neural Network

What Is It?

- Non-linear relationship between inputs and output
- Prediction more important than ease of explaining model
- Requires a lot of training data
- Users can specify the number of hidden layers, the number of hidden neurons, and associated activation functions for each layer
- Users can configure Input and Target Standardizations, Target Error, and Activation Functions



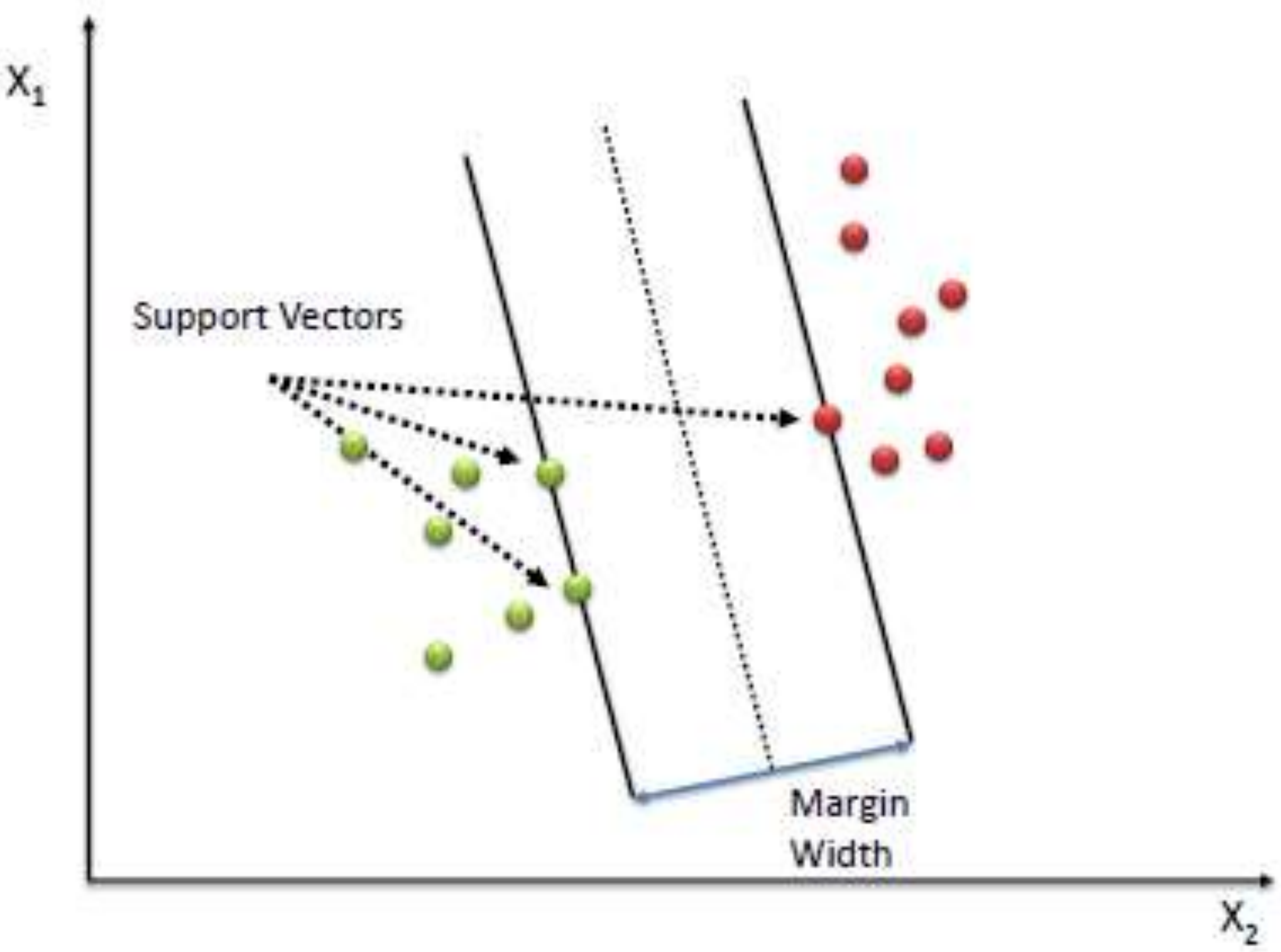
Many types…

- Feedforward Neural Network
- Radial Basis Function Neural Network
- Multilayer Perceptron
- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)
- Modular Neural Network.
- Sequence-To-Sequence Models

Support Vector Machines

What Is It?

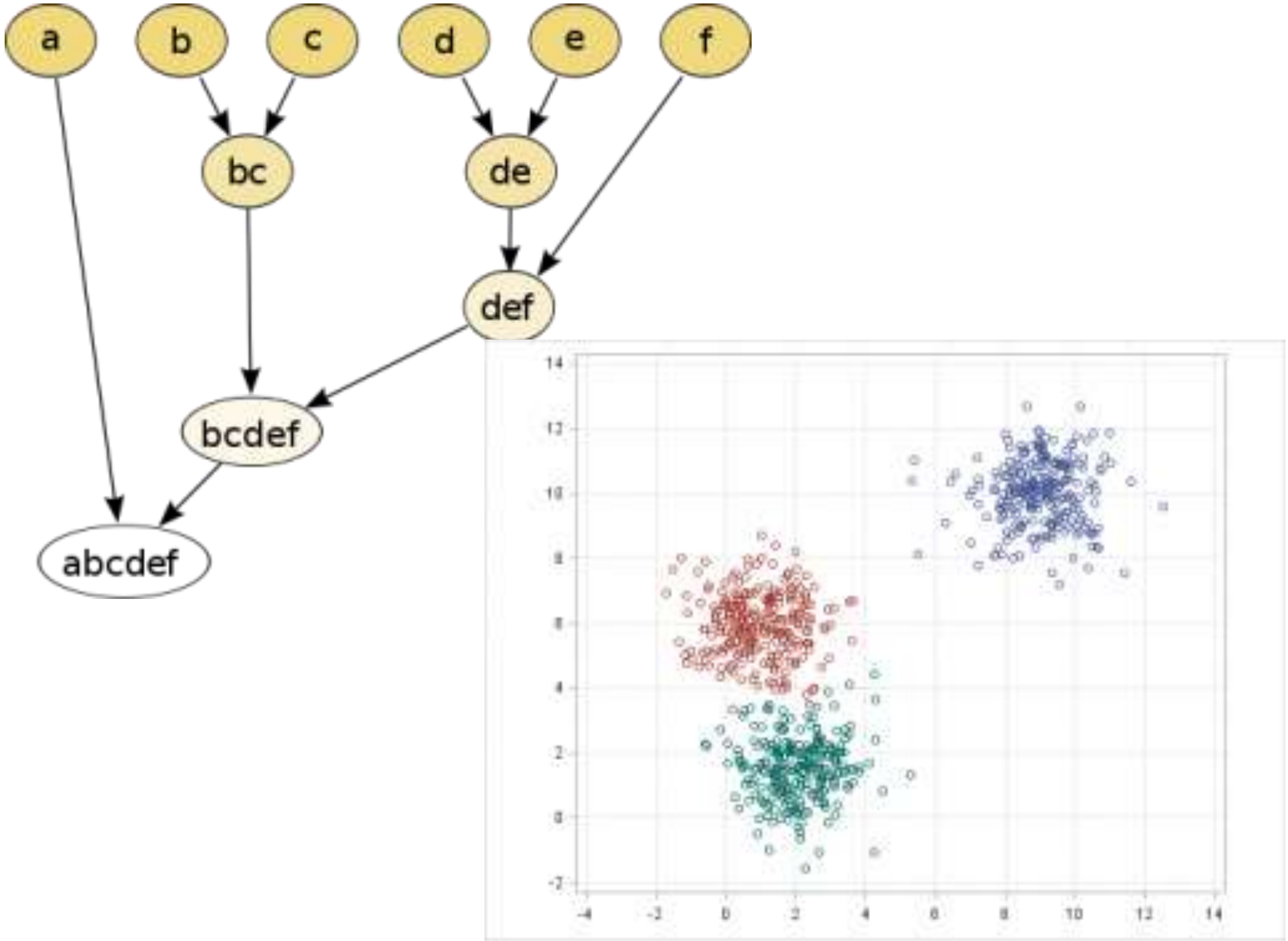
- Enables the creation of linear and nonlinear support vector machine models
- Constructs separating hyperplanes that maximize the margin between two classes
- The vectors (cases) that define the hyperplane are the support vectors
- Enables use of a variety of kernels: linear, polynomial, radia basis function, and sigmoid function. The node also provides interior point and active set optimization methods.



Clustering

What Is It?

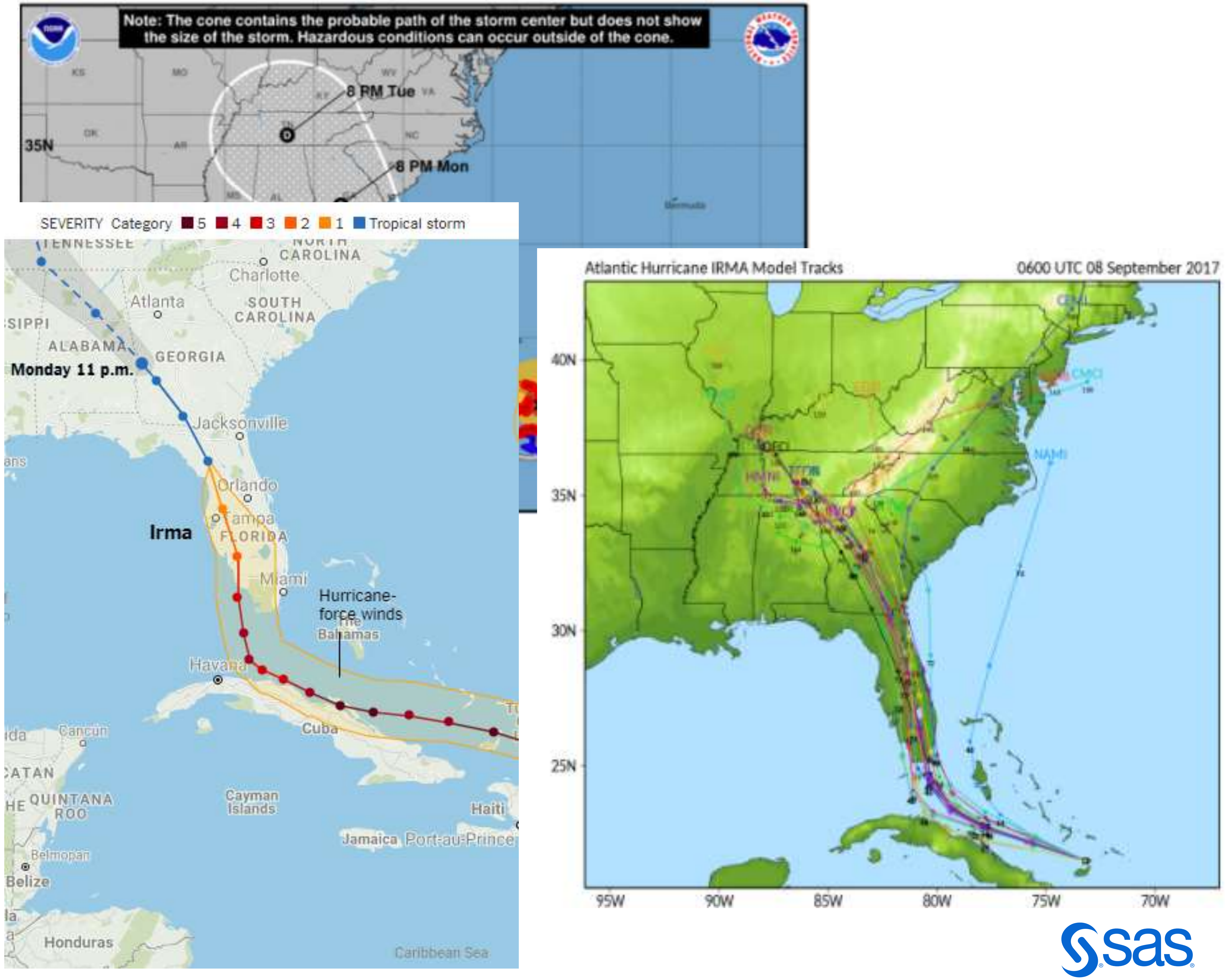
- Goal: The goal of clustering is to partition data into groups so that the observations within a group are as similar as possible to each other, and as dissimilar as possible to the observations in other groups.
- Many types - Hierarchical, k-means, SOM, etc..



Ensemble Modeling

What Is It?

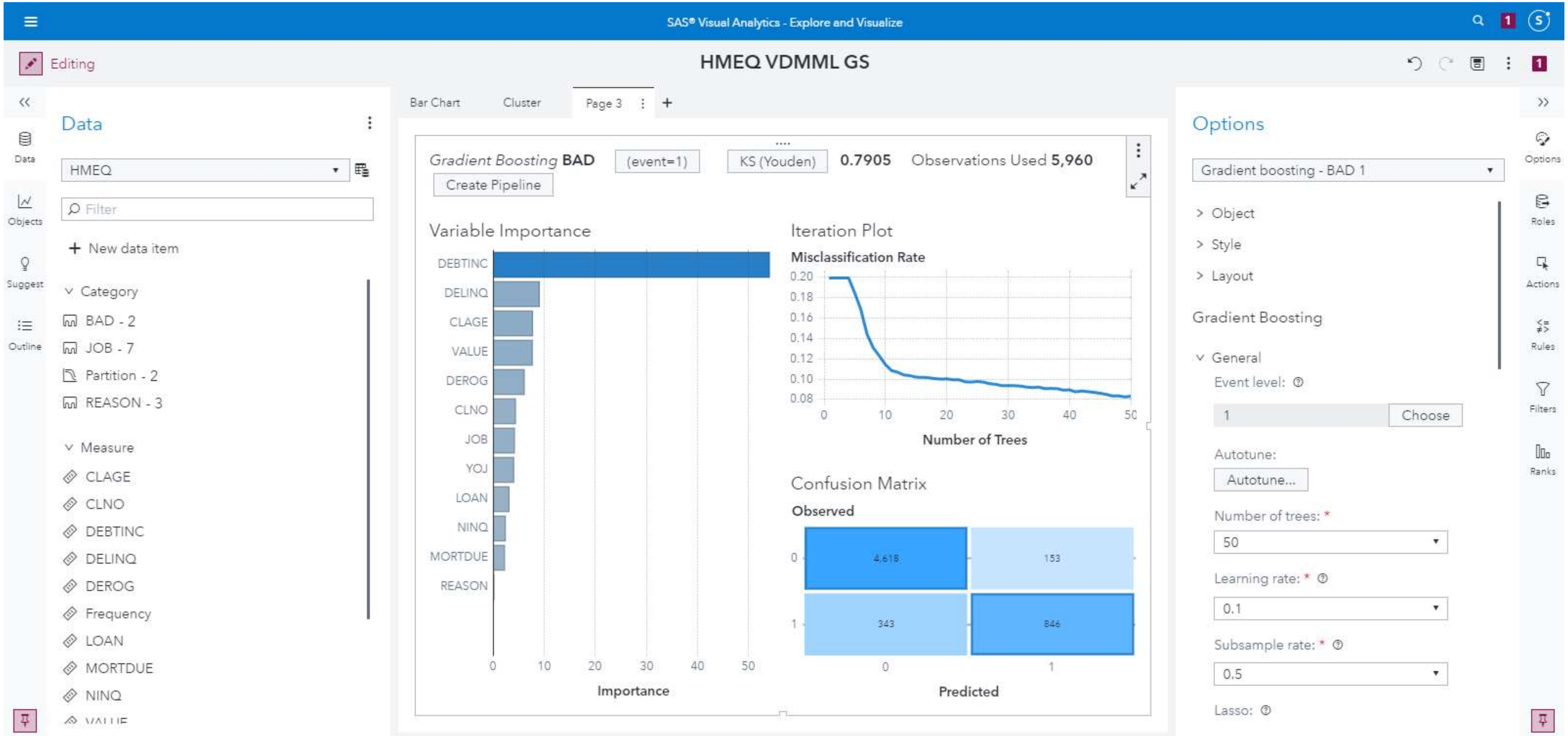
- Two or more predictive models combined to create a potentially more accurate model
- Works better when model predictions are uncorrelated
- Creates new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple predecessor models.
- 3 Methods
  - Average
  - Maximum
  - Voting



Machine Learning in SAS Viya

Interfaces

Building a Model from Scratch in the Visual Reporting Interface



HMEQ VDMML GS

Gradient Boosting **BAD** (event=1) KS (Youden) 0.7905 Observations Used **5,960**

Gradient boosting - BAD 1

Event level: 1

Autotune:

Number of trees: * 50

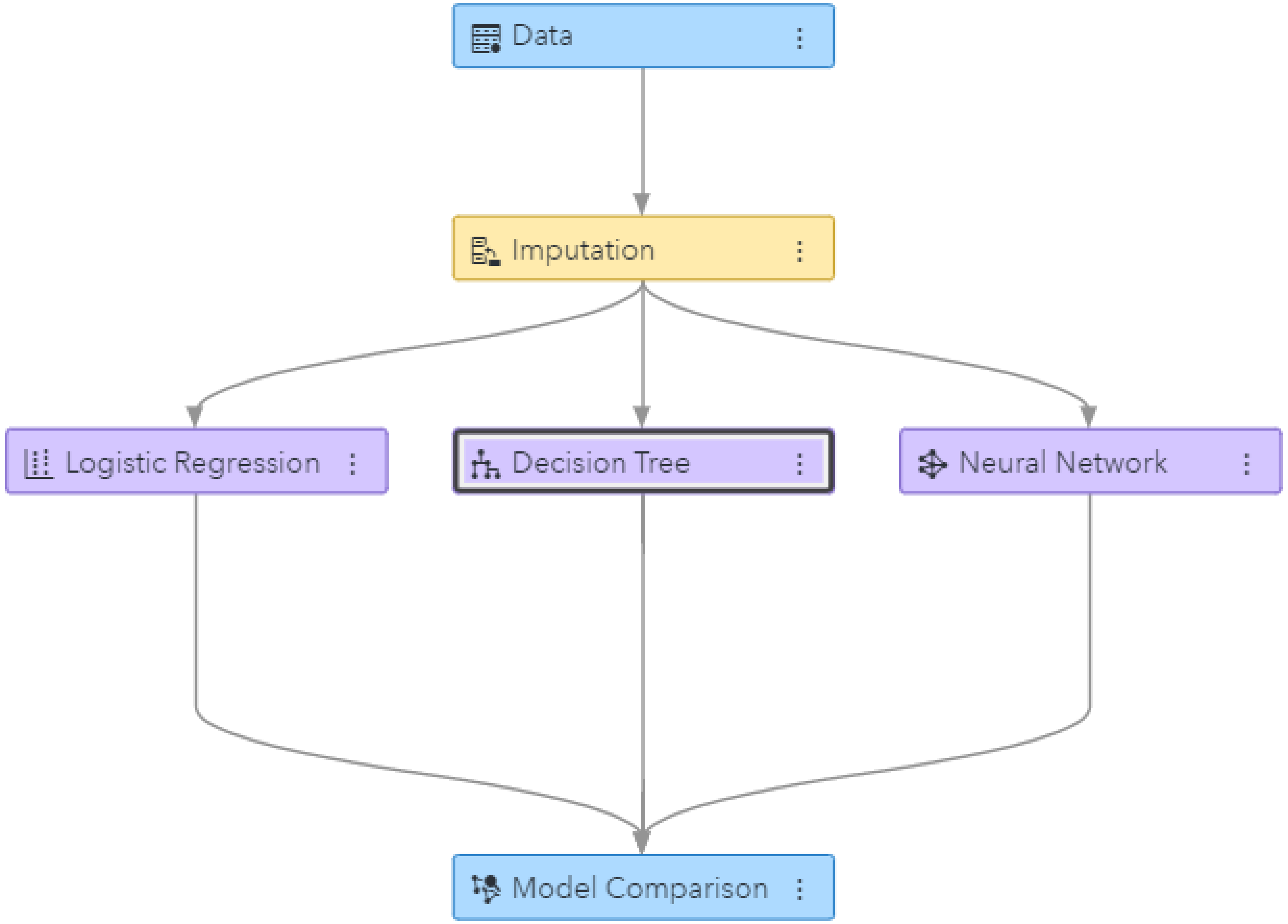
Learning rate: * 0.1

Subsample rate: * 0.5

Interfaces

Build Models Using Pipelines in Model Studio

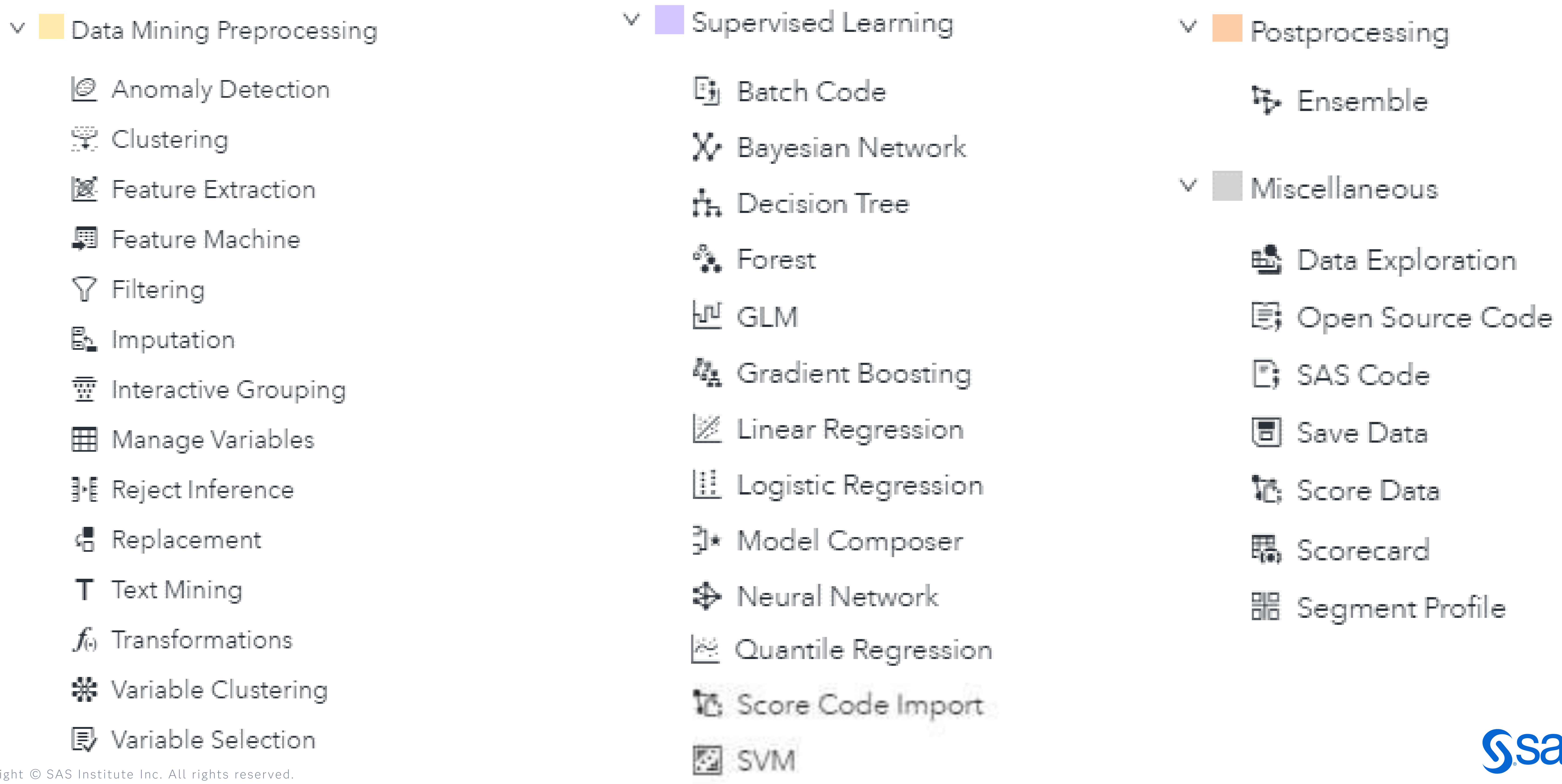
- Drag-and-drop pipelines including preprocessing and machine learning techniques
- Customizable and portable nodes and SAS best practice pipelines (Toolbox)
- Support for SAS coding (macro, data step, procs, batch Enterprise Miner) within pipelines
- Collaboration using the “Toolbox” – a collection of SAS Best Practice Pipelines, in addition to user-generated templates



Example Code for Pipeline

SAS® Visual Data Mining and Machine Learning

Pipelines



- Data Mining Preprocessing
  - Anomaly Detection
  - Clustering
  - Feature Extraction
  - Feature Machine
  - Filtering
  - Imputation
  - Interactive Grouping
  - Manage Variables
  - Reject Inference
  - Replacement
  - Text Mining
  - Transformations
  - Variable Clustering
  - Variable Selection

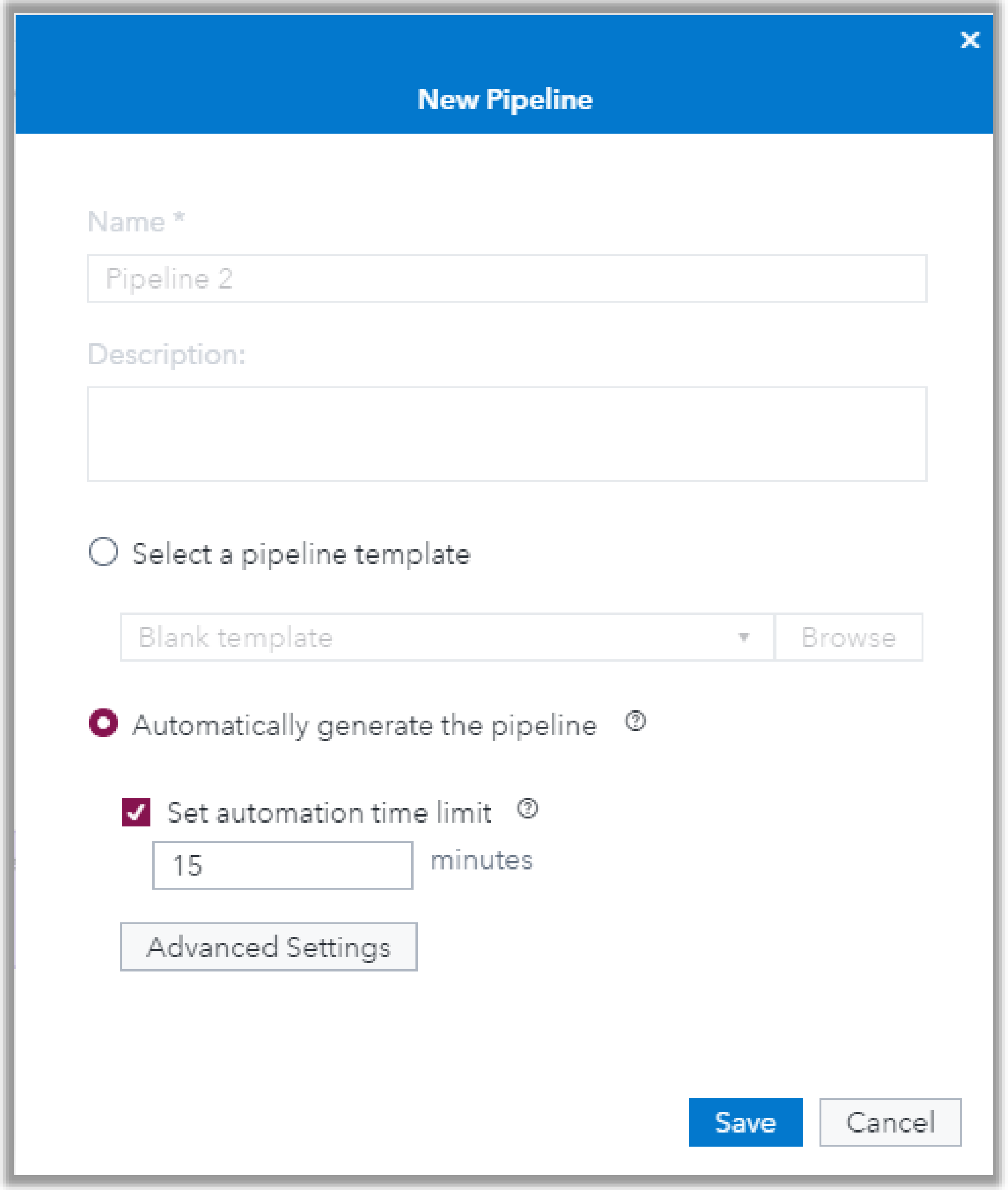
- Supervised Learning
  - Batch Code
  - Bayesian Network
  - Decision Tree
  - Forest
  - GLM
  - Gradient Boosting
  - Linear Regression
  - Logistic Regression
  - Model Composer
  - Neural Network
  - Quantile Regression
  - Score Code Import
  - SVM

- Postprocessing
  - Ensemble

- Miscellaneous
  - Data Exploration
  - Open Source Code
  - SAS Code
  - Save Data
  - Score Data
  - Scorecard
  - Segment Profile

Building Pipelines

Use prebuilt templates or automatically generate the pipeline



×

New Pipeline

Name *

Pipeline 2

Description:

○ Select a pipeline template

Blank template ▾ Browse

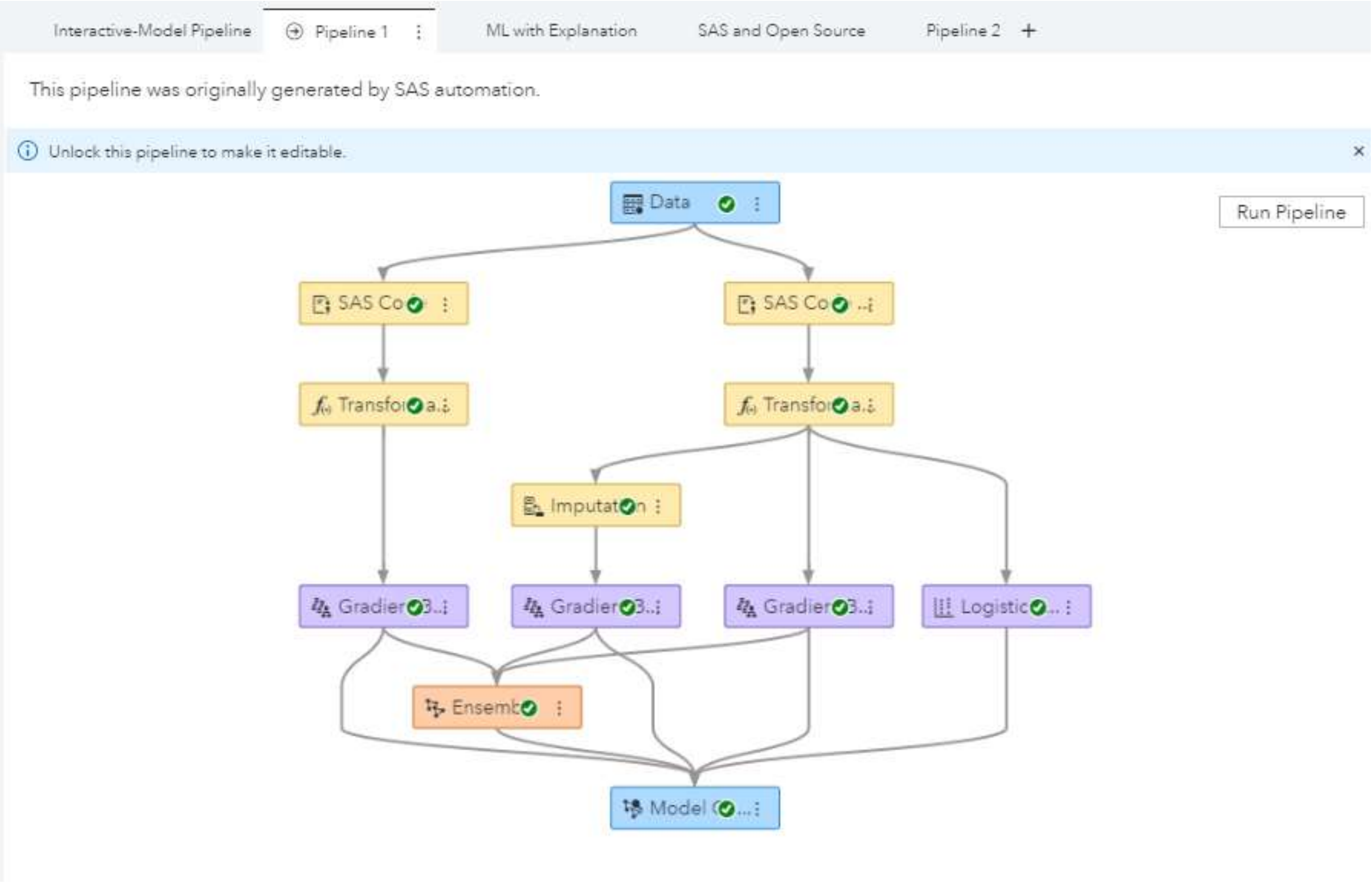
● Automatically generate the pipeline ⓘ

☑ Set automation time limit ⓘ

15 minutes

Advanced Settings

Save Cancel

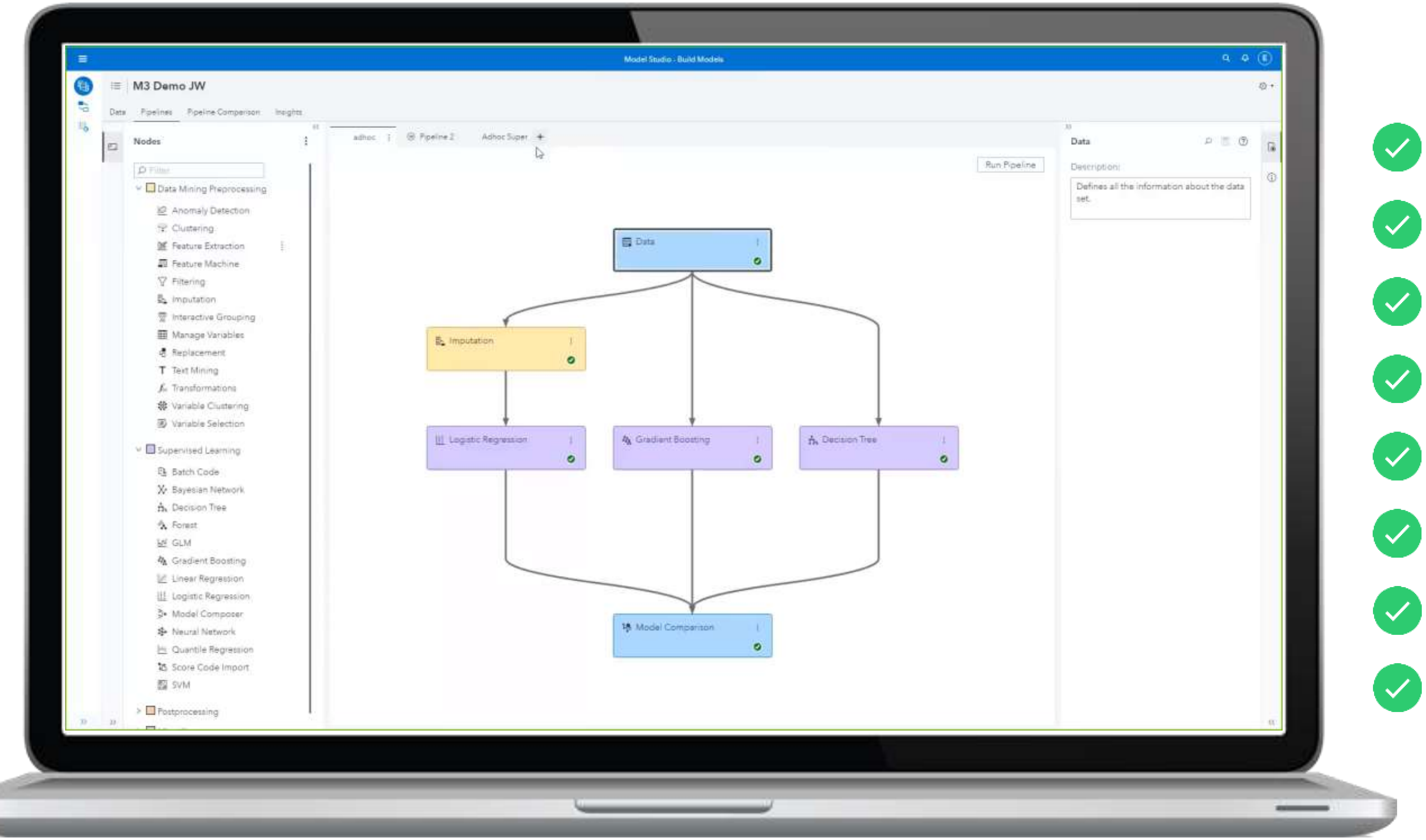


Interactive-Model Pipeline | Pipeline 1 | ML with Explanation | SAS and Open Source | Pipeline 2 +

This pipeline was originally generated by SAS automation.

ⓘ Unlock this pipeline to make it editable. ×

Automated Pipelines

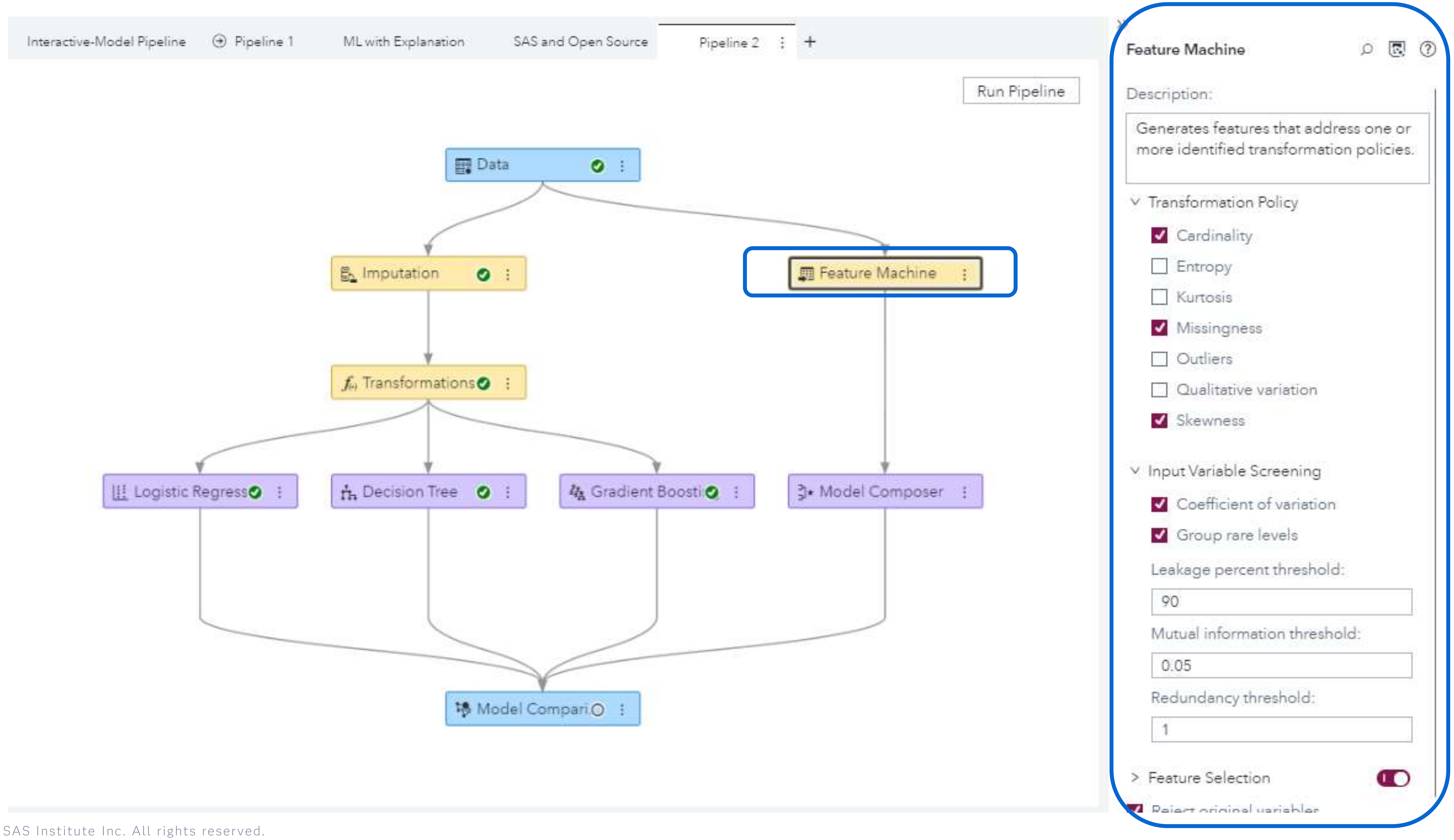


Defines all the information about the data set.

- ✓ Repository of best practice pipelines
- ✓ Models by SAS or by end-user
- ✓ Dynamically reads thru data
- ✓ Fixes data quality issues w/ ML
- ✓ Performs Data transformations
- ✓ Recommends & builds models
- ✓ Optimizes across models
- ✓ Fully editable, no black-box

Automatically Create Features

Feature Machine Node



Feature Machine

Description:

Generates features that address one or more identified transformation policies.

Transformation Policy

- ☑ Cardinality
- □ Entropy
- □ Kurtosis
- ☑ Missingness
- □ Outliers
- □ Qualitative variation
- ☑ Skewness

Input Variable Screening

- ☑ Coefficient of variation
- ☑ Group rare levels

Leakage percent threshold: 90

Mutual information threshold: 0.05

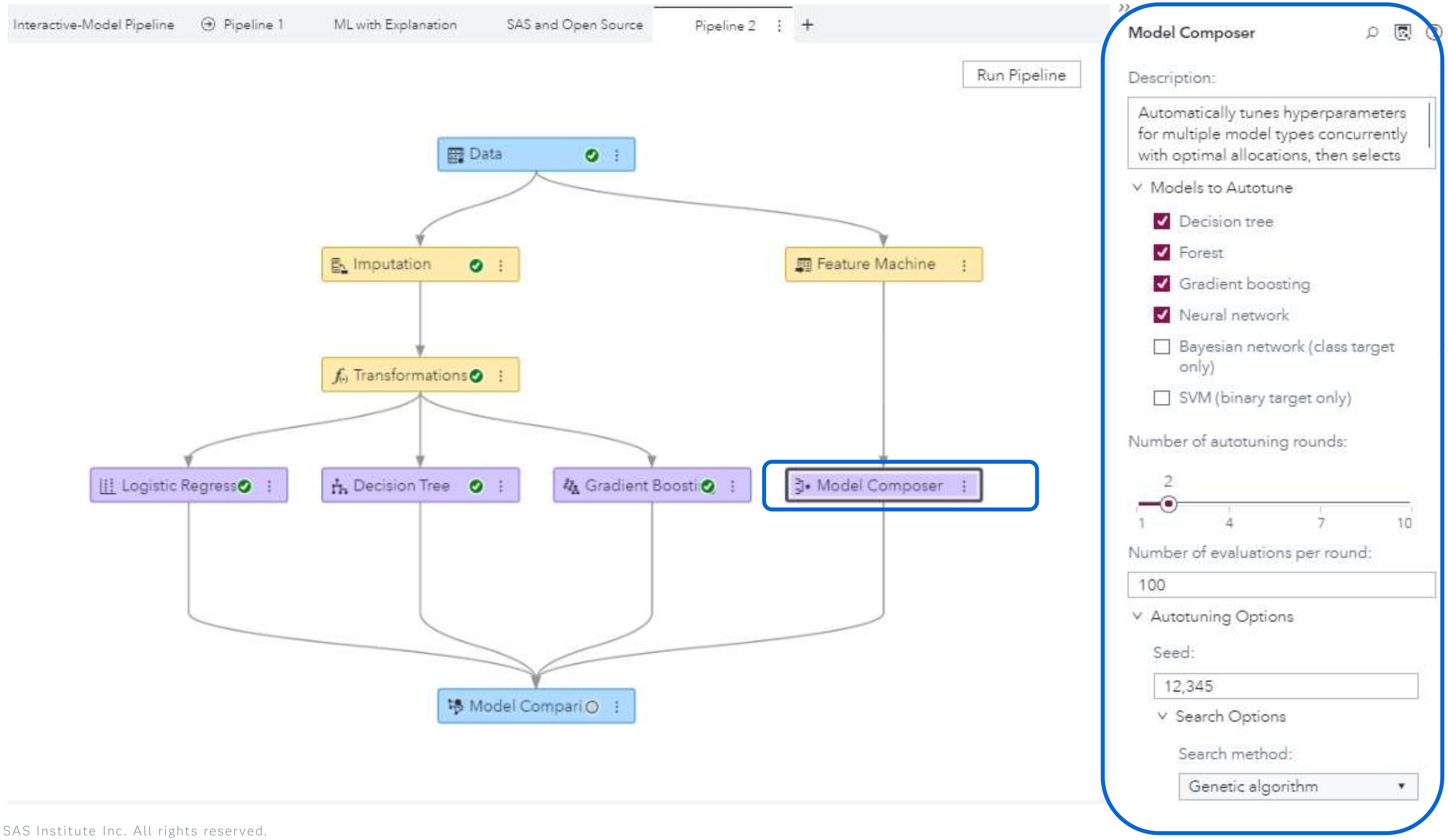
Redundancy threshold: 1

Feature Selection ●

☑ Reject original variables

Automatically Tune Hyperparameters for Multiple Model Types

Model Composer Node



Model Composer

Description:

Automatically tunes hyperparameters for multiple model types concurrently with optimal allocations, then selects

Models to Autotune

☑ Decision tree

☑ Forest

☑ Gradient boosting

☑ Neural network

□ Bayesian network (class target only)

□ SVM (binary target only)

Number of autotuning rounds: 2

1 4 7 10

Number of evaluations per round: 100

Autotuning Options

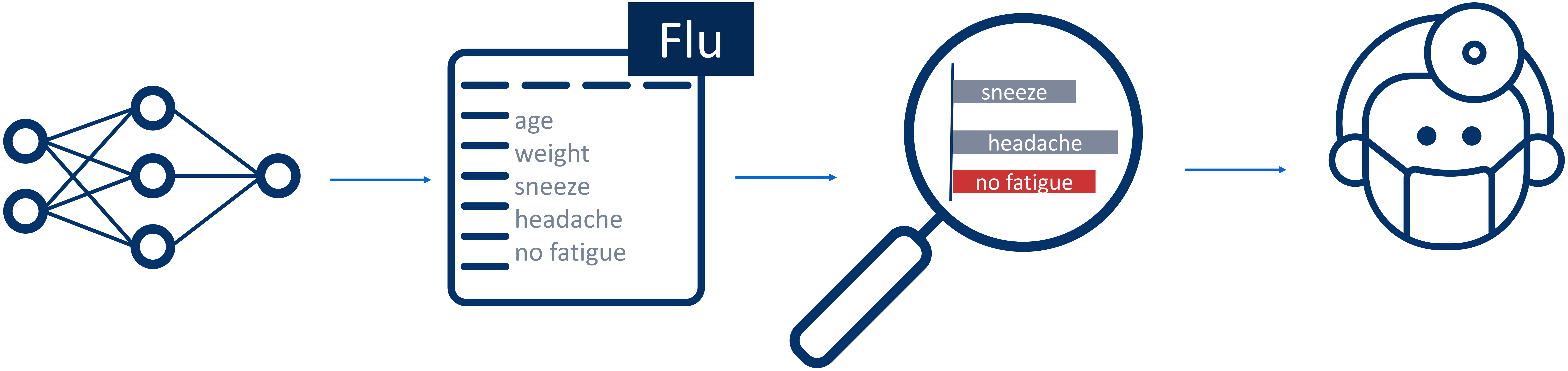
Seed: 12,345

Search Options

Search method: Genetic algorithm

Interpretable Machine Learning

Why is it important?



Model

Data & Prediction

Explanation

Human Decision

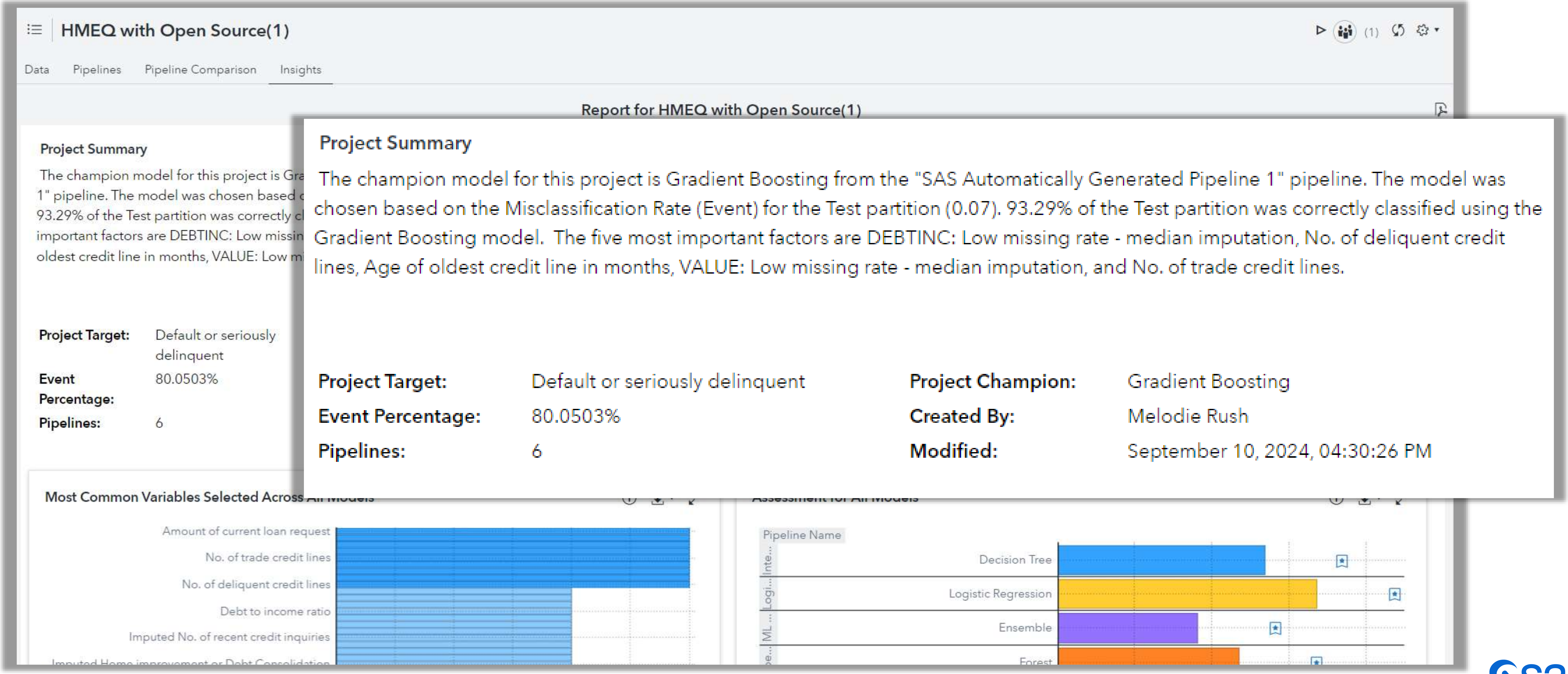
Interpretable Machine Learning

Popular Approaches

1. Variable Importance
2. Partial Dependency Plots
3. Individual Conditional Expectation (ICE)
4. Local Interpretable Model-Agnostic Explanation (LIME)
5. SHapley Additive exPlanations (SHAP)

Automated Insights & Interpretability

Description in simple language



HMEQ with Open Source(1)
(1)
Data
Pipelines
Pipeline Comparison
Insights
Report for HMEQ with Open Source(1)
Most Common Variables Selected Across All Models
Amount of current loan request
No. of trade credit lines
No. of deliquent credit lines
Debt to income ratio
Imputed No. of recent credit inquiries
Imputed Home improvement or Debt Consolidation
Assessment for All Models
Pipeline Name
Decision Tree
Logistic Regression
Ensemble
Forest

Project Summary

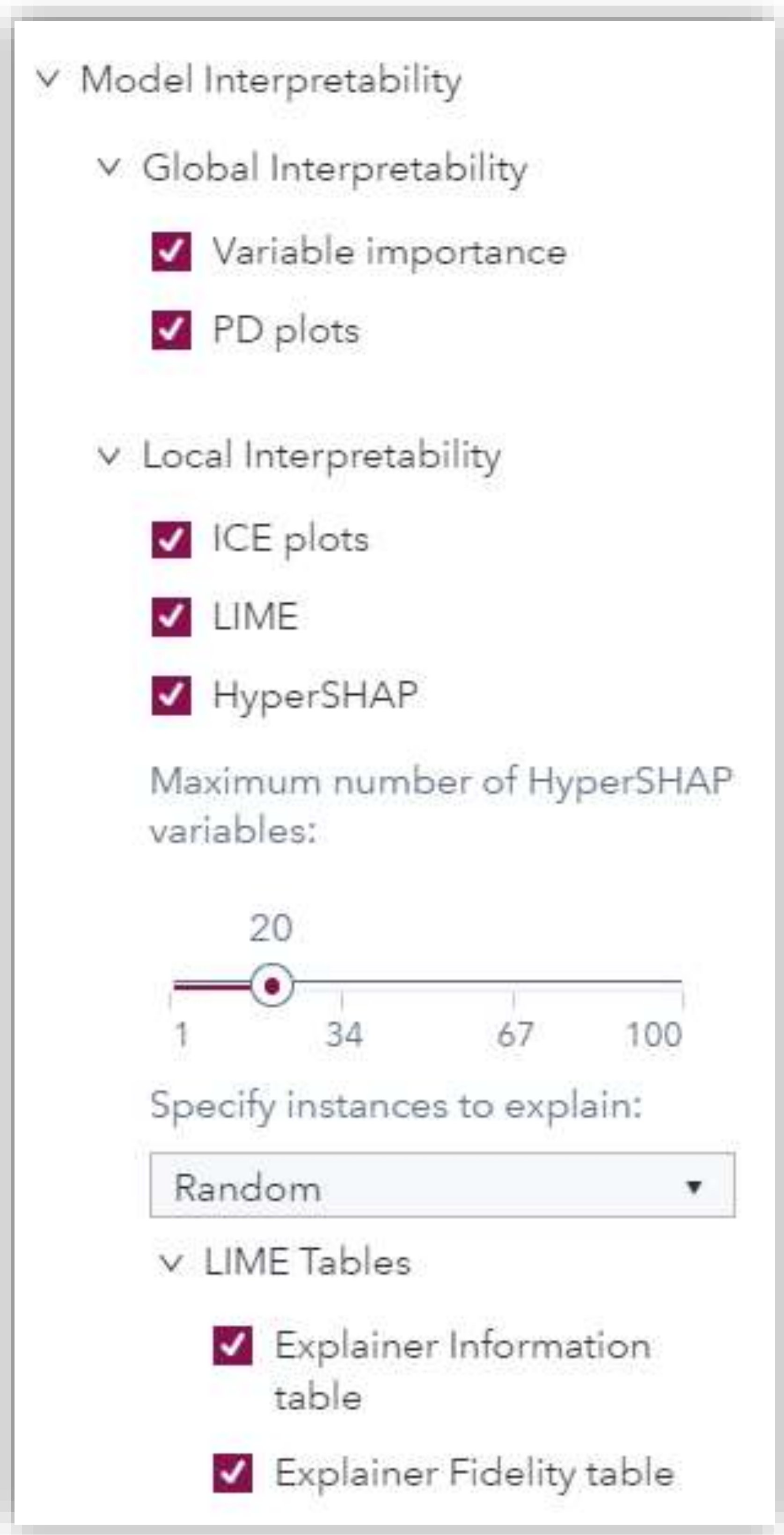
The champion model for this project is Gradient Boosting from the "SAS Automatically Generated Pipeline 1" pipeline. The model was chosen based on the Misclassification Rate (Event) for the Test partition (0.07). 93.29% of the Test partition was correctly classified using the Gradient Boosting model. The five most important factors are DEBTINC: Low missing rate - median imputation, No. of deliquent credit lines, Age of oldest credit line in months, VALUE: Low missing rate - median imputation, and No. of trade credit lines.

Project Target:	Default or seriously delinquent	**Project Champion:**	Gradient Boosting
Event Percentage:	80.0503%	**Created By:**	Melodie Rush
Pipelines:	6	**Modified:**	September 10, 2024, 04:30:26 PM

Automated Insights & Interpretability

Model Interpretability Charts

- Variable Importance Plots and Rankings
- Partial Dependence (PD) Plots
- LIME (Local Interpretable Model-agnostic Explanations)
- ICE (Individual Conditional Expectation) Plots
- Kernel SHAP Method (Shapley Values)



Model Interpretability

Global Interpretability

☑ Variable importance

☑ PD plots

Local Interpretability

☑ ICE plots

☑ LIME

☑ HyperSHAP

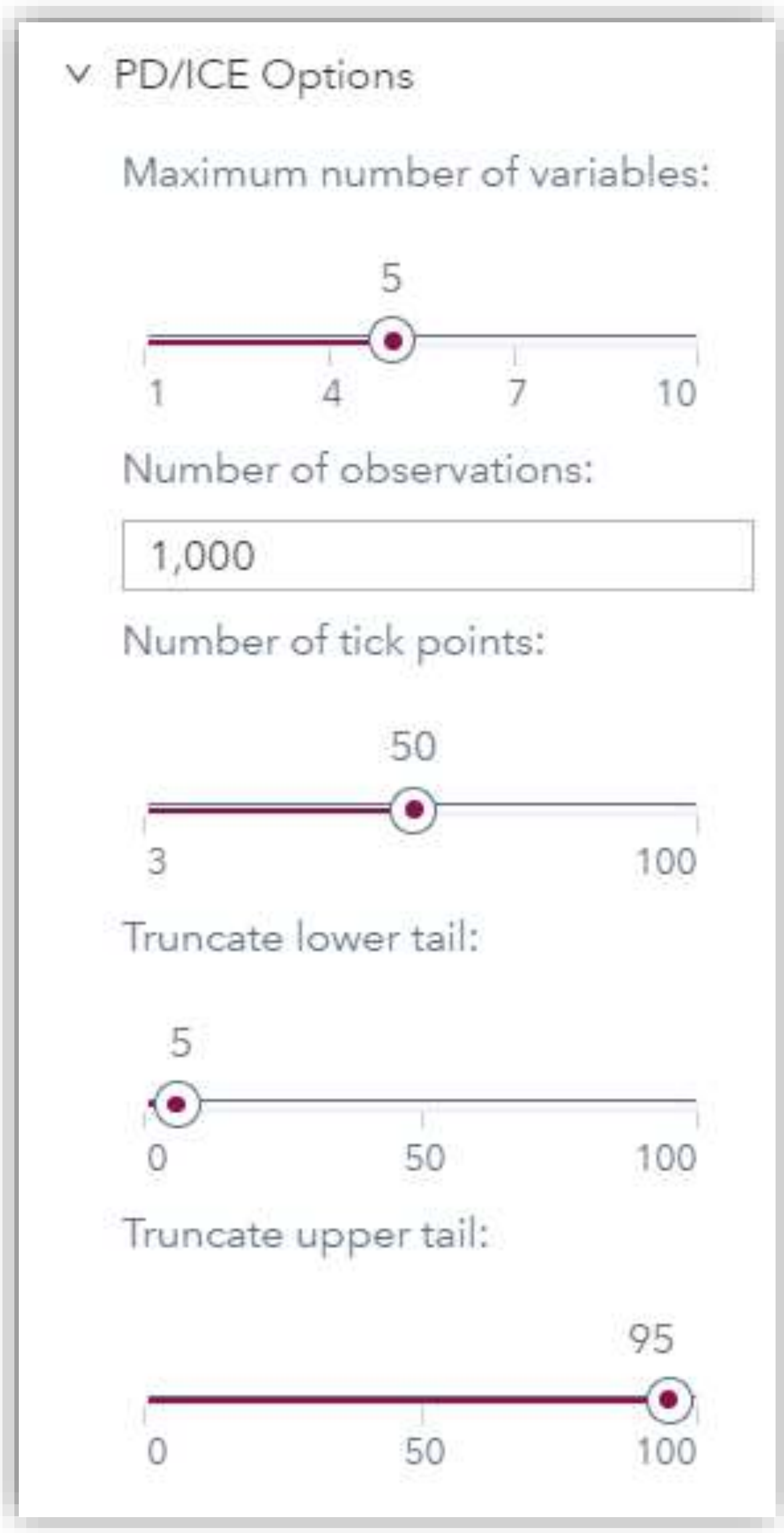
Maximum number of HyperSHAP variables: 20

Specify instances to explain: Random ▾

LIME Tables

☑ Explainer Information table

☑ Explainer Fidelity table



PD/ICE Options

Maximum number of variables: 5

Number of observations: 1,000

Number of tick points: 50

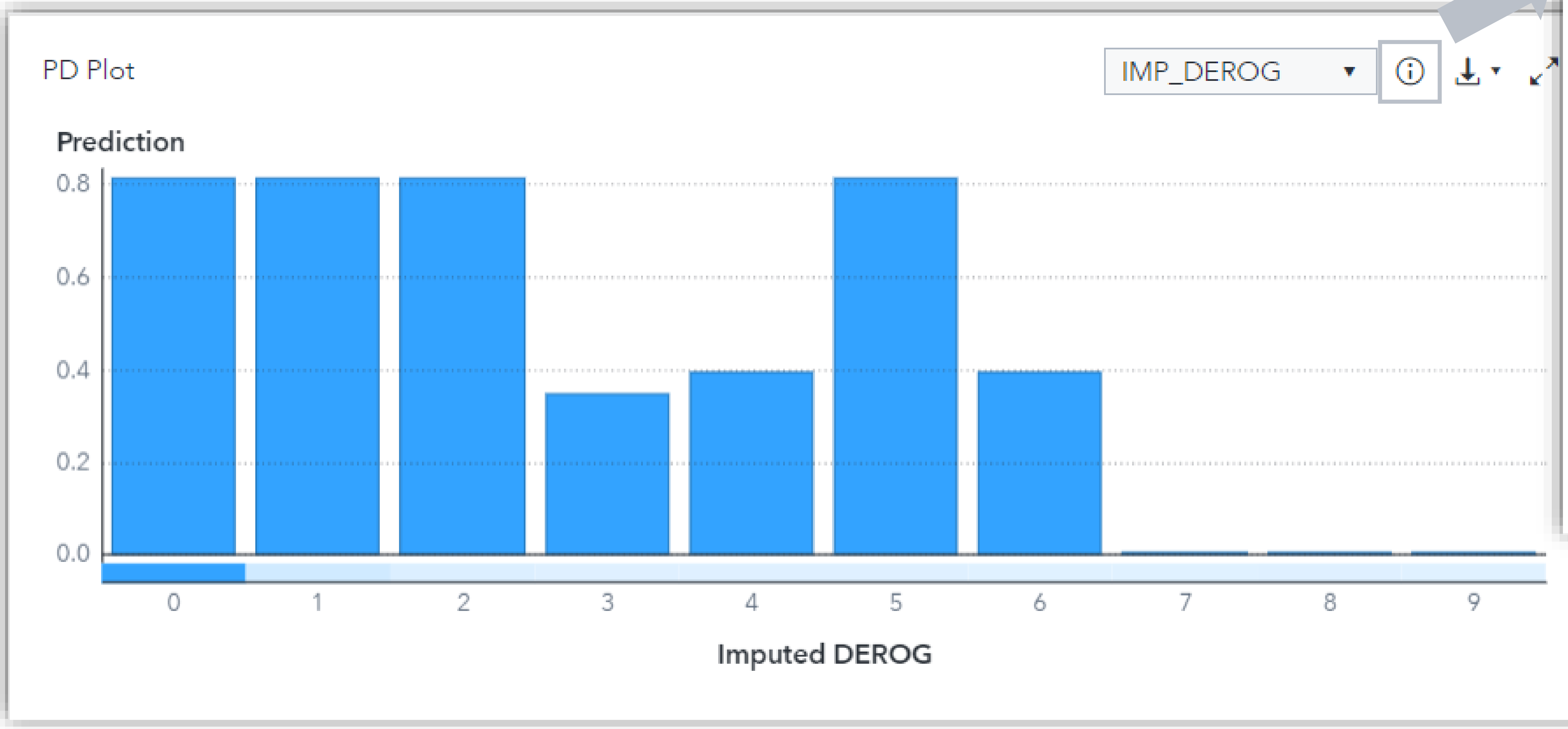
Truncate lower tail: 5

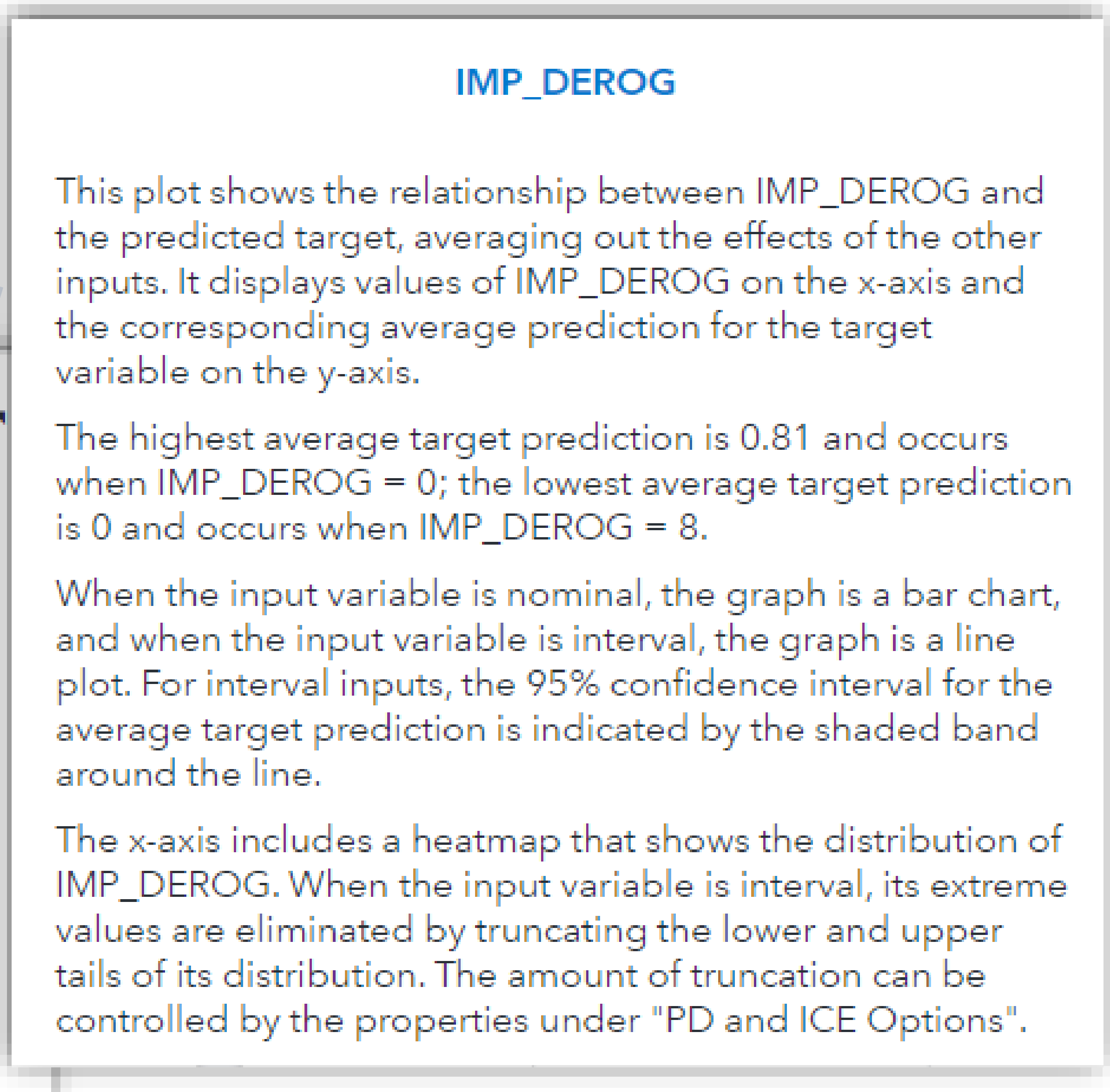
Truncate upper tail: 95

Automated Insights & Interpretability

Model Interpretability Charts

Each interpretability chart has insights included





IMP_DEROG

This plot shows the relationship between IMP_DEROG and the predicted target, averaging out the effects of the other inputs. It displays values of IMP_DEROG on the x-axis and the corresponding average prediction for the target variable on the y-axis.

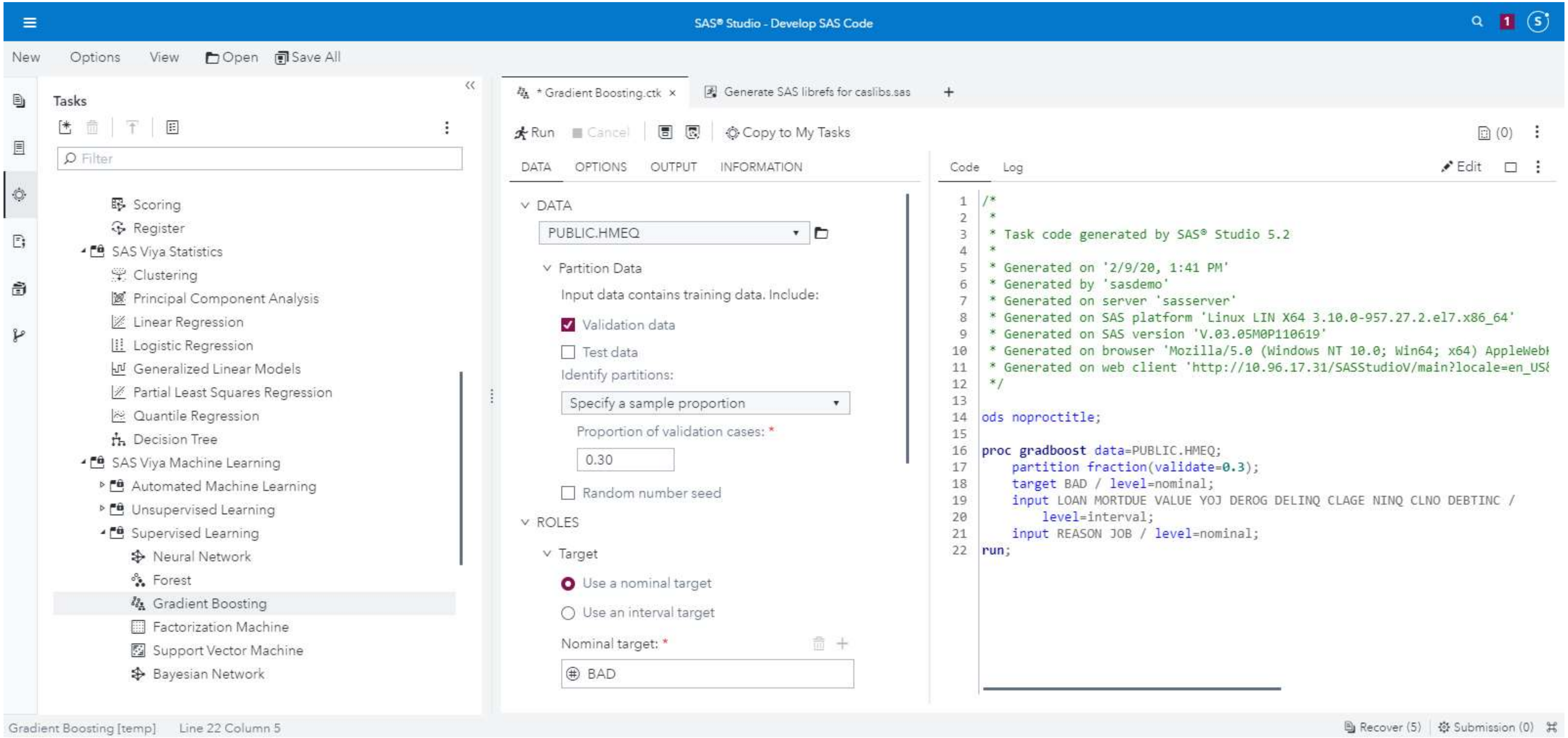
The highest average target prediction is 0.81 and occurs when IMP_DEROG = 0; the lowest average target prediction is 0 and occurs when IMP_DEROG = 8.

When the input variable is nominal, the graph is a bar chart, and when the input variable is interval, the graph is a line plot. For interval inputs, the 95% confidence interval for the average target prediction is indicated by the shaded band around the line.

The x-axis includes a heatmap that shows the distribution of IMP_DEROG. When the input variable is interval, its extreme values are eliminated by truncating the lower and upper tails of its distribution. The amount of truncation can be controlled by the properties under "PD and ICE Options".

Interfaces

Building a Model Using SAS Studio Tasks



DATA

PUBLIC.HMEQ

Partition Data

Input data contains training data. Include:

☑ Validation data

☐ Test data

Identify partitions:

Specify a sample proportion

Proportion of validation cases: *

0.30

☐ Random number seed

ROLES

Target

● Use a nominal target

○ Use an interval target

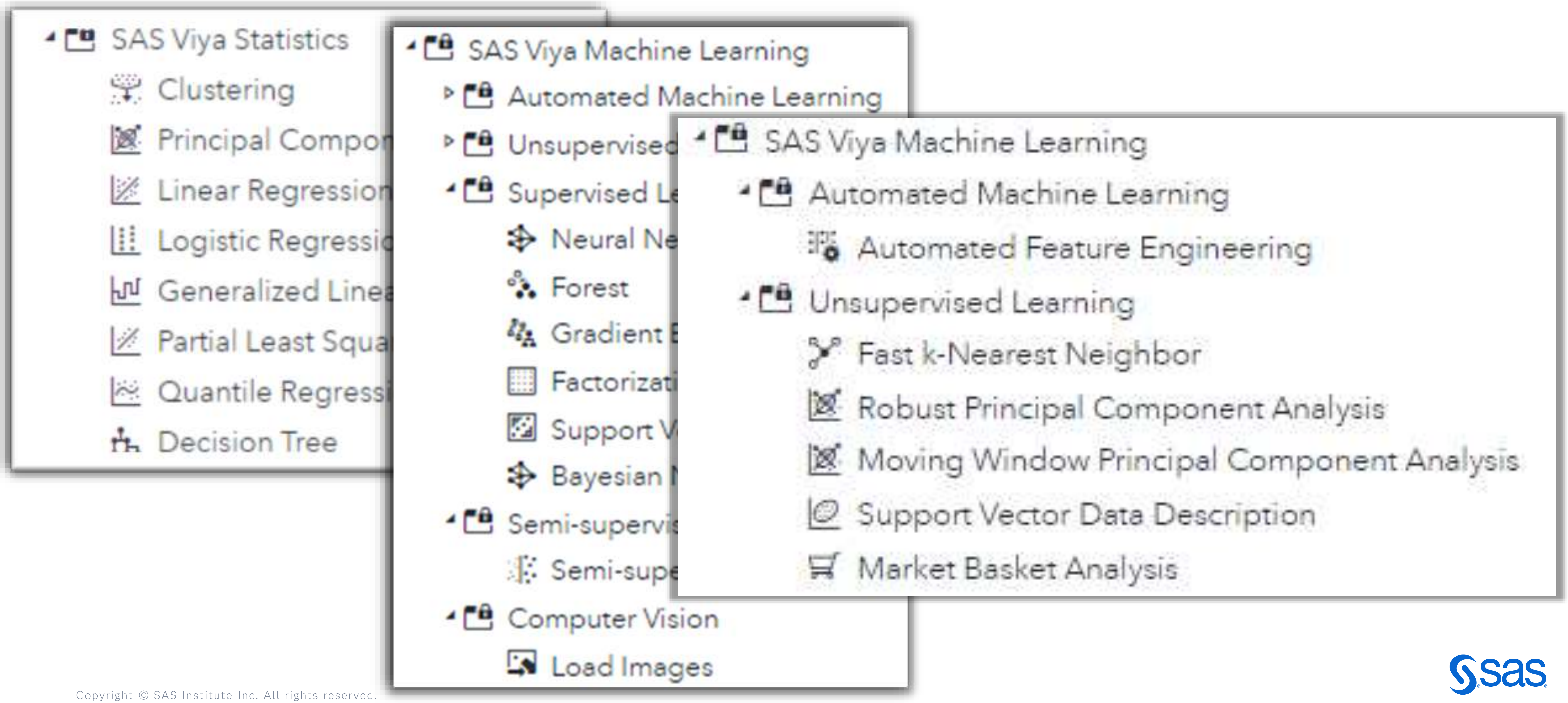
Nominal target: *

BAD

```
/*
 *
 * Task code generated by SAS® Studio 5.2
 *
 * Generated on '2/9/20, 1:41 PM'
 * Generated by 'sasdemo'
 * Generated on server 'sasserver'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-957.27.2.el7.x86_64'
 * Generated on SAS version 'V.03.05M0P110619'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebK
 * Generated on web client 'http://10.96.17.31/SASStudioV/main?locale=en_US&
 */

ods noproctitle;

proc gradboost data=PUBLIC.HMEQ;
    partition fraction(validate=0.3);
    target BAD / level=nominal;
    input LOAN MORTDUE VALUE YOJ DEROG DELINQ CLAGE NINQ CLNO DEBTINC /
        level=interval;
    input REASON JOB / level=nominal;
run;
```

SAS Visual Data Mining and Machine Learning

Programming Tasks in SAS Studio



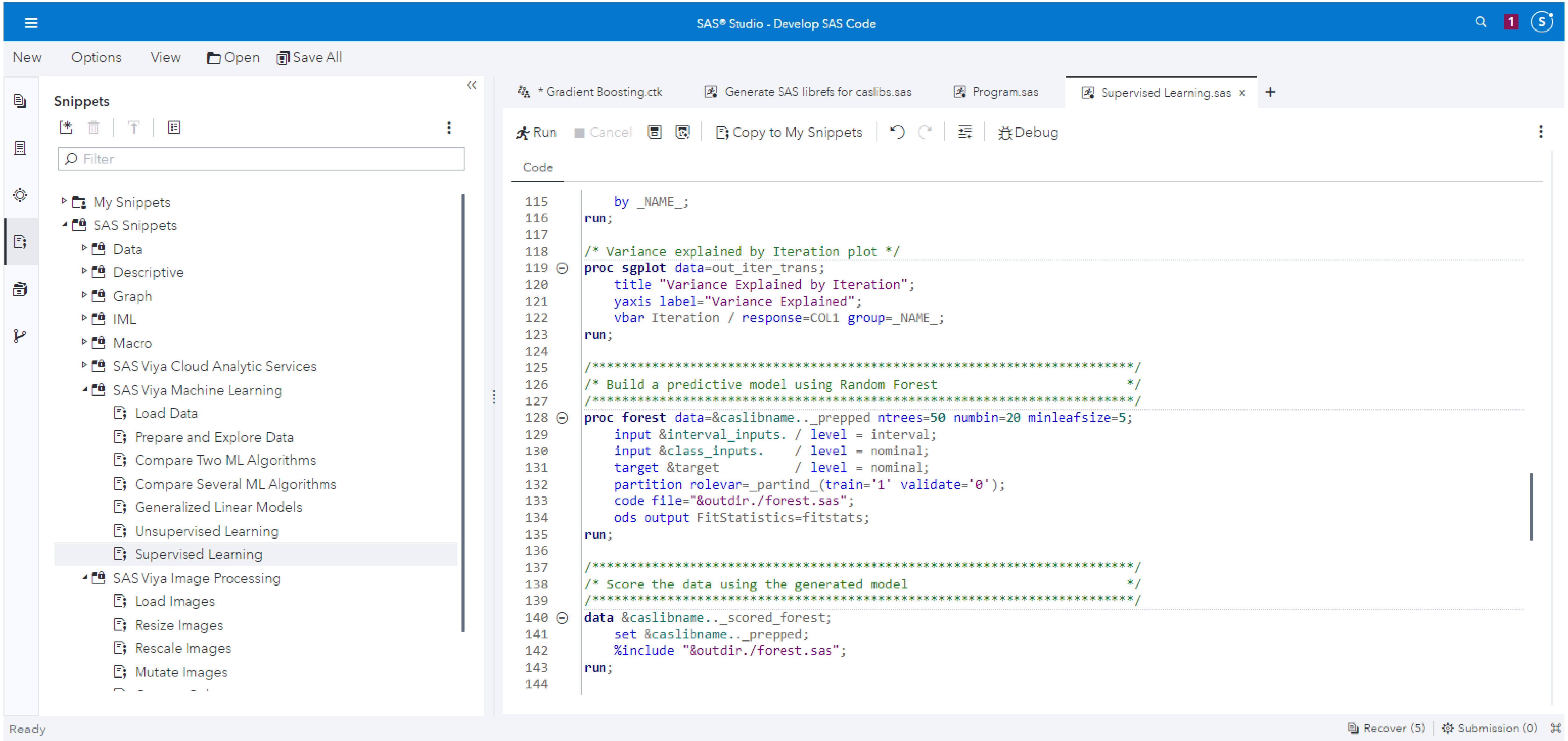
- SAS Viya Statistics
  - Clustering
  - Principal Compon
  - Linear Regression
  - Logistic Regressio
  - Generalized Linea
  - Partial Least Squa
  - Quantile Regressi
  - Decision Tree

- SAS Viya Machine Learning
  - Automated Machine Learning
  - Unsupervised
  - Supervised Le
    - Neural Ne
    - Forest
    - Gradient B
    - Factorizati
    - Support V
    - Bayesian N
  - Semi-supervis
    - Semi-supe
  - Computer Vision
    - Load Images

- SAS Viya Machine Learning
  - Automated Machine Learning
    - Automated Feature Engineering
  - Unsupervised Learning
    - Fast k-Nearest Neighbor
    - Robust Principal Component Analysis
    - Moving Window Principal Component Analysis
    - Support Vector Data Description
    - Market Basket Analysis

Interfaces

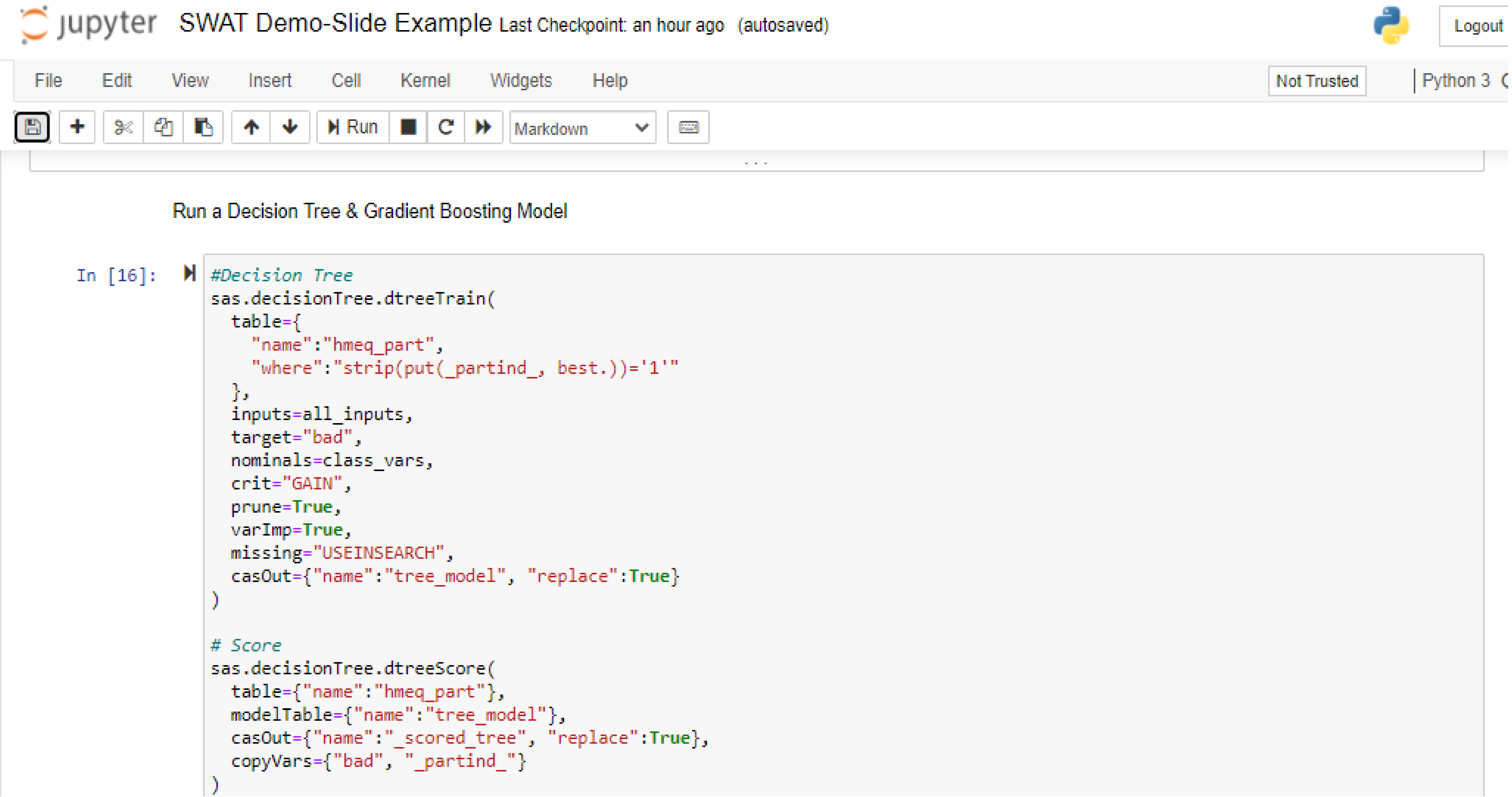
Building a Model Using SAS Studio Snippets



```
    by _NAME_;
run;

/* Variance explained by Iteration plot */
proc sgplot data=out_iter_trans;
    title "Variance Explained by Iteration";
    yaxis label="Variance Explained";
    vbar Iteration / response=COL1 group=_NAME_;
run;

/************************************************************************/
/* Build a predictive model using Random Forest                         */
/************************************************************************/
proc forest data=&caslibname.._prepped ntrees=50 numbin=20 minleafsize=5;
    input &interval_inputs. / level = interval;
    input &class_inputs.    / level = nominal;
    target &target          / level = nominal;
    partition rolevar=_partind_(train='1' validate='0');
    code file="&outdir./forest.sas";
    ods output FitStatistics=fitstats;
run;

/************************************************************************/
/* Score the data using the generated model                             */
/************************************************************************/
data &caslibname.._scored_forest;
    set &caslibname.._prepped;
    %include "&outdir./forest.sas";
run;
```

Interfaces

Building a Model Using Open Source



Run a Decision Tree & Gradient Boosting Model

In [16]:

```
#Decision Tree
sas.decisionTree.dtreeTrain(
  table={
    "name":"hmeq_part",
    "where":"strip(put(_partind_, best.))='1'"
  },
  inputs=all_inputs,
  target="bad",
  nominals=class_vars,
  crit="GAIN",
  prune=True,
  varImp=True,
  missing="USEINSEARCH",
  casOut={"name":"tree_model", "replace":True}
)

# Score
sas.decisionTree.dtreeScore(
  table={"name":"hmeq_part"},
  modelTable={"name":"tree_model"},
  casOut={"name":"_scored_tree", "replace":True},
  copyVars={"bad", "_partind_"}
)
```

Review

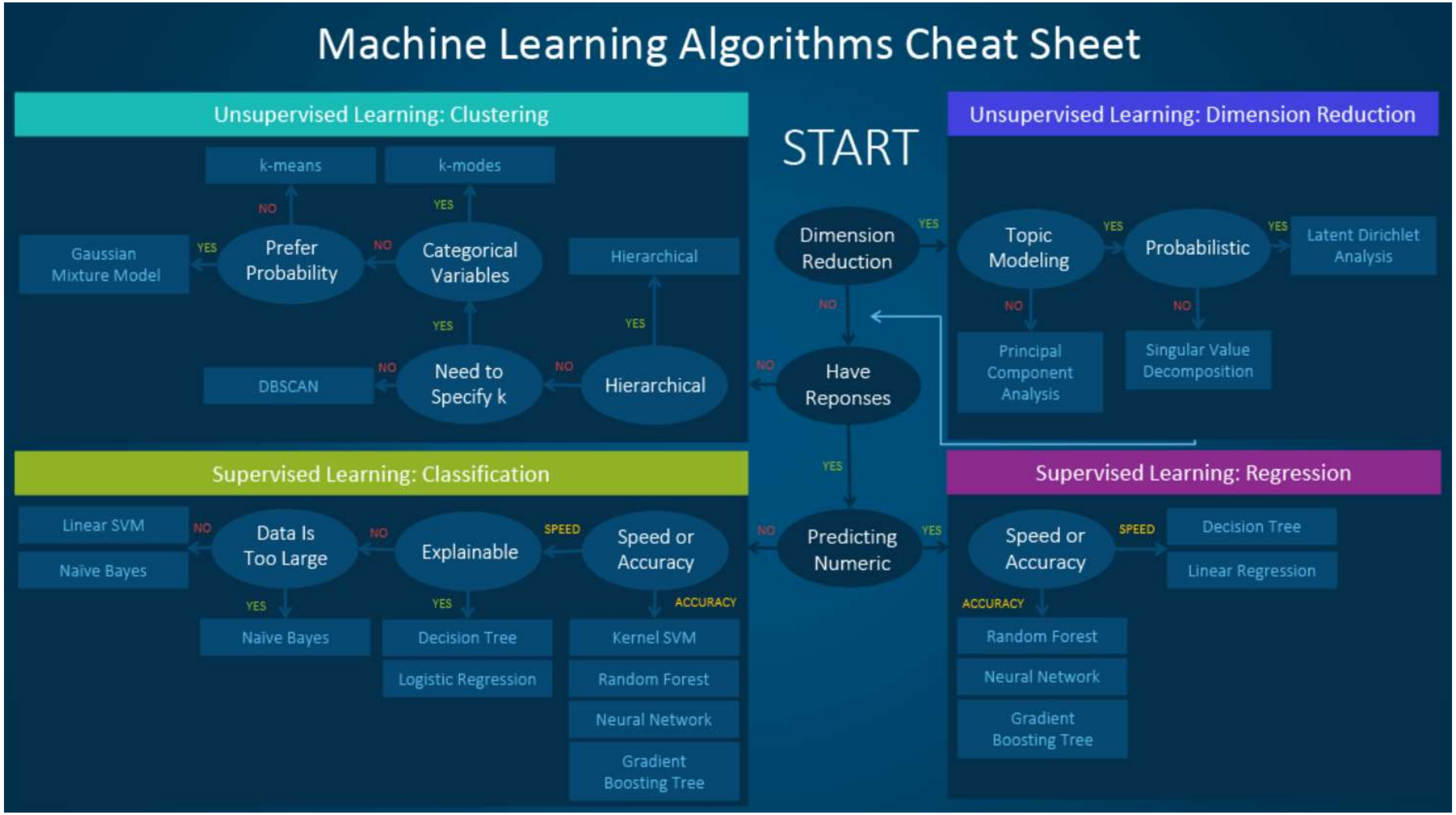
What we covered today

- What is Machine Learning?
- Machine Learning Terminology
- Intro to ML Modeling Algorithms
- Machine Learning in SAS Viya

Resources

Where to learn more

Machine Learning Algorithms Cheat Sheet



[Access Here]

Recommended Resources

An Overview of SAS® Visual Data Mining and Machine Learning on SAS® Viya
https://support.sas.com/resources/papers/proceedings17/SAS1492-2017.pdf

Video - Automated Machine Learning at Scale
http://www.sas.com/en_us/webinars/automated-machine-learning-scale.html

Machine learning - what it is and why it matters (reading)
http://www.sas.com/en_us/insights/analytics/machine-learning.html

Live web and classroom training - Big Data, Data Mining, and Machine Learning
Big Data course

SAS Tutorial

Videos

How to Choose a Machine Learning Algorithm
https://youtu.be/-oZcf0QEzYM

Transforming variables in SAS
https://communities.sas.com/t5/SAS-Data-Mining-and-Machine/New-video-Transforming-Variables-in-SAS/m-p/710687#M8553

Thank you for your time and attention!

Introduction to Machine Learning

Connect with me:
LinkedIn: https://www.linkedin.com/in/melodierush
Twitter: @Melodie_Rush