BANKRUPTCY CLASSIFICATION

Using SAS Enterprise Miner

MGMT 571

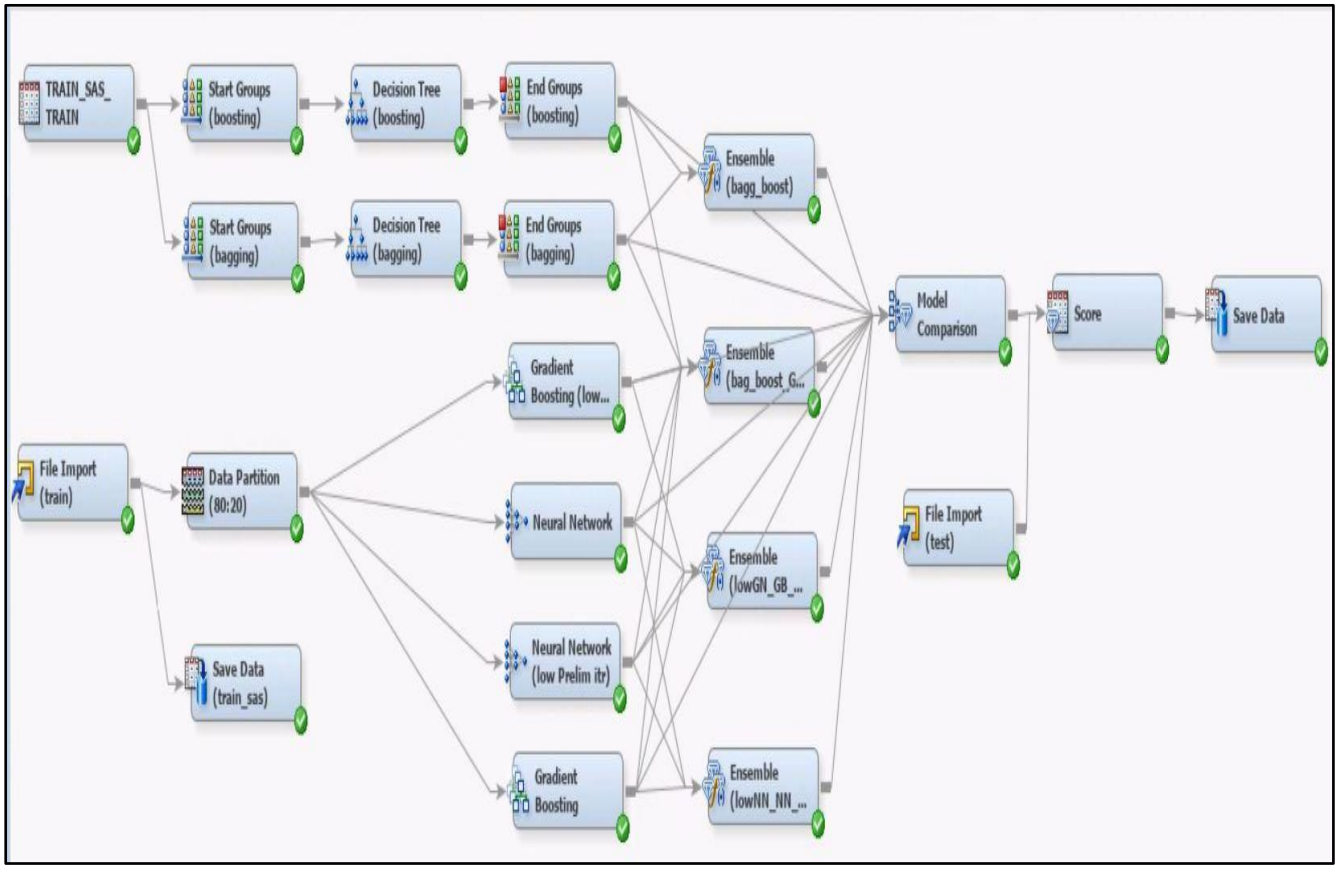
Data Mining Final Project

Team: Data-cated Miners

Kalp Shah, Prerak Patel, Mayank Jha

PURDUE
KRANNERT
SCHOOL OF MANAGEMENT

1. SAS EM Model



PURDUE
KRANNERT
SCHOOL OF MANAGEMENT

2. Description

An **Ensemble** model of bagging Decision-Tree, Boosting Decision-Tree, Gradient Boosting, Gradient Boosting with lower shrinkage, Neural Network and Neural Network with lower optimization parameter gave best Validation ROC of **0.999**.

The training data used for Decision trees were in sas7dat format, so as to facilitate bagging and boosting with Start-group and end-group nodes.

Decision Tree (Boosting):

Property	Value
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	Most correlated branch
Use Input Once	No
Maximum Branch	30
Maximum Depth	15
Minimum Categorical Size	5
Node	
Leaf Size	2
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	Yes
Number of Subsets	5
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes

Decision Tree (Bagging):

Property	Value
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Most correlated branch
Use Input Once	No
Maximum Branch	30
Maximum Depth	15
Minimum Categorical Size	2
Node	
Leaf Size	2
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	Yes
Number of Subsets	5
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	

PURDUE
KRANNERT
SCHOOL OF MANAGEMENT

Gradient Boosting (Low Shrinkage):

Property	Value
General	
Node ID	Boost
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Series Options	
N Iterations	5000
Seed	12345
Shrinkage	0.1
Train Proportion	100
Splitting Rule	
Huber M-Regression	0.9
Maximum Branch	12
Maximum Depth	30
Minimum Categorical Size	2
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Most correlated branch
Performance	Disk
Node	
Leaf Fraction	0.1
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	

Gradient Boosting:

Property	Value
General	
Node ID	Boost3
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Series Options	
N Iterations	5000
Seed	12345
Shrinkage	0.2125
Train Proportion	100
Splitting Rule	
Huber M-Regression	0.9
Maximum Branch	12
Maximum Depth	30
Minimum Categorical Size	2
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Most correlated branch
Performance	Disk
Node	
Leaf Fraction	0.1
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	

Neural network (Low Preliminary Iterations):

Property	Value
Training Technique	Trust-Region
Maximum Iterations	1000
Maximum Time	7 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
Maximum Momentum	1.75
Tilt	0.0
Preliminary Training	
Enable	Yes
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

1.2

Neural Network:

Property	Value
Training Technique	Trust-Region
Maximum Iterations	1000
Maximum Time	7 Hours
Nonlinear Options	
Use Defaults	No
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
Maximum Momentum	1.75
Tilt	0.0
Preliminary Training	
Enable	Yes
Number of Runs	10
Maximum Iterations	100
Maximum Time	7 Hours

3. Model Performance

Selected Model	Model Node	Model Description	Selection Criterion: Valid: Roc Index
Y	Ensmbl	Ensemble (bag_boost_GB_lowGB_NN_lowNN)	0.999
	Ensmbl4	Ensemble (lowGN_GB_lowNN_NN)	0.973
	Ensmbl3	Ensemble (lowNN_NN_GB)	0.968
	Neural2	Neural Network	0.938
	Neural	Neural Network (low Prelim itr)	0.938
	Boost	Gradient Boosting	0.938
	Boost3	Gradient Boosting (low Shrinkage)	0.929