

Modeling Rare Events

Over/Undersampling, Priors, Decision Weights

Undersampling/ Oversampling and Prior Probabilities

• • •

Can be accounted for automatically in SAS EM

Undersampling and Prior Probabilities

- Say you have a rare event as target (<10% of data)
 - Fraud
 - Catastrophic failure
 - 10%+ single day change in value of stock market index
- May have trouble modelling because a model is accurate for classifying everything as nonevent!
- Potential Solution: Create a biased sample

Undersampling and Prior Probabilities

- Potential Solution: Create a biased sample
 - **Undersample:** under-represent common events in training data.
 - Keep all rare events and only a fraction of common events
 - Ratio of Common:Rare events is up for debate.
 - 70:30 ought to be fine.
 - 50:50 is sometimes encouraged.
 - **Oversample:**
 - replicate the rare events in training.
 - do this *after* the training/validation split so don't have the same observation in both training and validation set!
 - OR, use a hybrid technique like **SMOTE** (Chawla, 2002) that creates new data points *like* the rare events (not exact replicates)

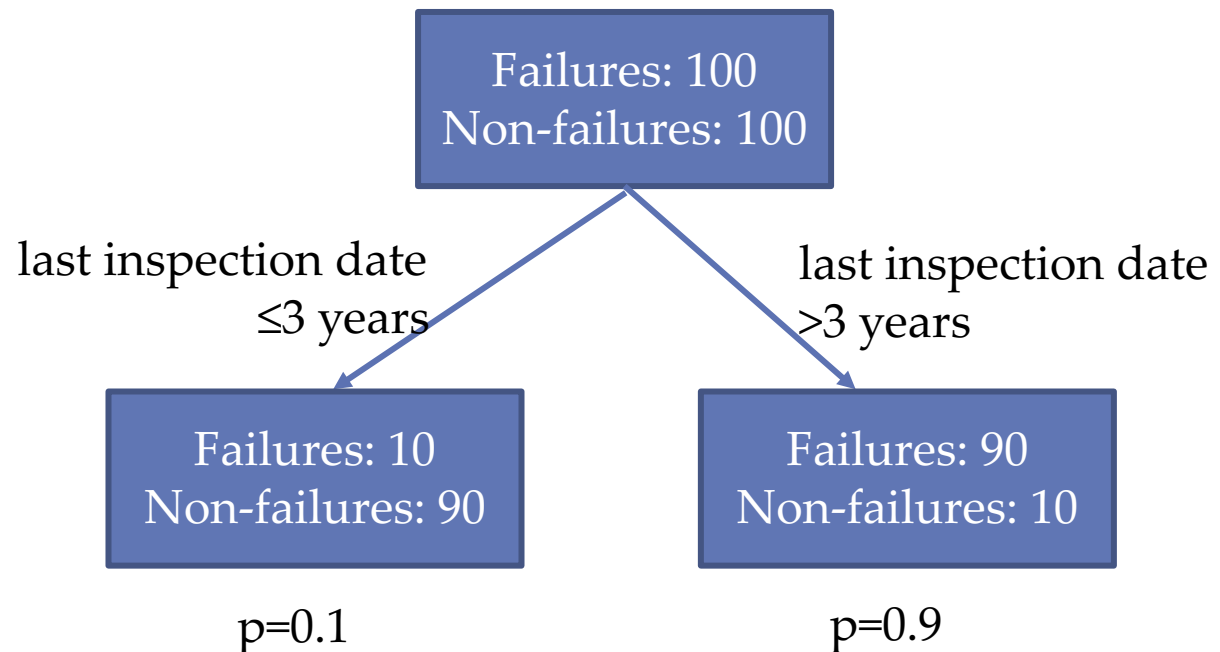
Undersampling and Prior Probabilities

- Models provide **posterior probabilities** for events.
- The accuracy of the posterior probabilities rely on a representative sample.
- If we bias our sample, must adjust the posterior probabilities to account for this.

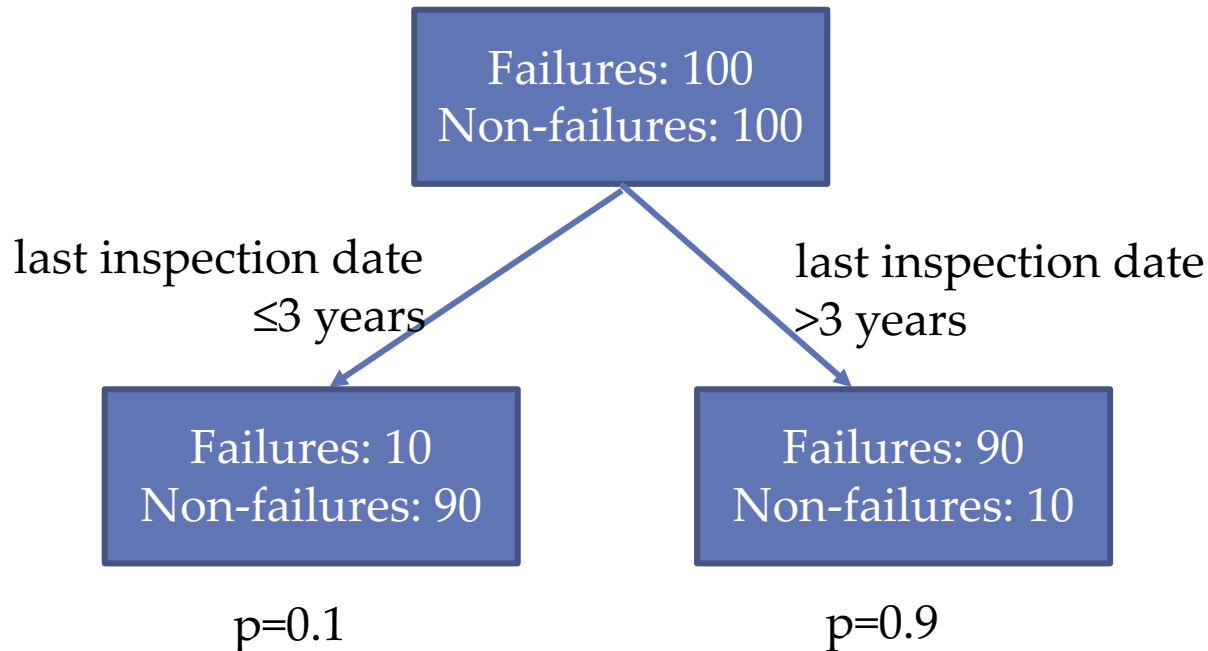
Why Adjustment is Necessary

Predict voting machine failure. Only 100 voting machines failed out of 10,000.

Undersample. Dataset has 100 failures and 100 non-failures.



Why Adjustment is Necessary



Does a new machine with last inspection date > 3 years really have a 90% probability of failing?

Why Adjustment is Necessary

- We'd have to go back to the data to answer this question.
- Assuming the 100 non-failures chosen were random, representative sample, we expect inspection date to be ≤ 3 years 90% of the time.
- That is 8,910 non-failing machines with inspection date ≤ 3 years.
- Similarly, 10% of non-failures have expect inspection date >3 years ago. This is 990 machines.

	≤ 3 years	>3 years
Failures	10	90
Nonfailures	8910	990

$P(\text{Failure} \mid \text{last inspection date} > 3 \text{ years})$
 $90/(90+990) = 8\%$
(Still failing at 8 times the rate of
recently inspected machines)

Summary: Adjusting for Undersampling

- Let $l = l_1, l_2, \dots, l_L$ be the levels of the target variable
- Let $i = 1, 2, \dots, n$ index the observations in the data
- Let $OldPost(i, l)$ be the posterior probability from the model on oversampled data
- Let $OldPrior(l)$ be the proportion of target level in the oversampled data
- Let $Prior(l)$ be the correct proportion of target level in true population

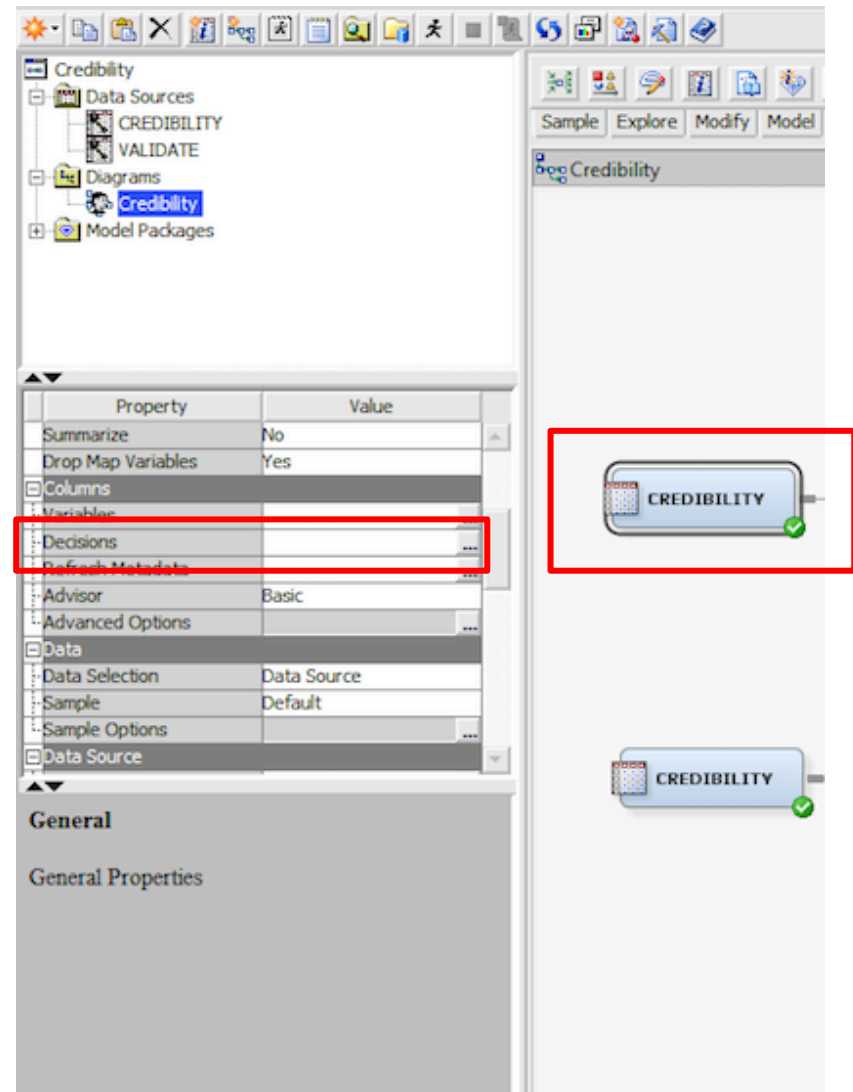
$$NewPost(i, l) = \frac{OldPost(i, l) \frac{Prior(l)}{OldPrior(l)}}{\sum_{j=1}^L OldPost(i, l_j) \frac{Prior(l_j)}{OldPrior(l_j)}}$$

Entering Priors and Decision Weights into SAS EM

...

Entering Priors into SAS EM

- Priors are also adjusted in the “decisions” on a dataset panel.
- Click “Build” when first opening the prompt, then click priors tab.



Only Some Output Uses the Prior Information

- In SAS EM, accounting for priors *has no effect on*:
 - Growing decision trees
 - Misclassification Rate (The cutoff probability is still 0.5 by default)
- Priors *do affect*:
 - *Pruning* decision trees
 - Once we account for a prior, a given split may not have a reasonable gain
- Net Effects:
 - Increasing a prior probability increases the posterior probability
 - Decreasing a prior decreases the posterior probability
 - Changing prior will have more noticeable effect if the original posterior is near 0.5 than if it is near 0 or 1.

Oversampled Data with No Priors

The screenshot shows the Enterprise Miner interface. The 'Exported Data - No Priors' window displays a table with the following data:

Port	Table	Role	Data Exists
TRAIN	EMWS1.Tree_TRAIN	Train	Yes
VALIDATE	EMWS1.Tree_VALIDATE	Validate	No
TEST	EMWS1.Tree_TEST	Test	No
SCORE	EMWS1.Tree_SCORE	Score	No
TRANSACTION	EMWS1.Tree_TRANSACTION	Transaction	No
TREE	EMWS1.Tree_EMTree	Tree	Yes

The 'Train' section of the 'Priors and Decisions' window shows the following properties:

Property	Value
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
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

The 'EMWS1.Tree_TRAIN' window shows the following data:

Node	Predicted: Creditability=bad	Predicted: Creditability=good	Unadjusted P: Creditability=bad
1	0.89473684210526	0.10526315789473	0.89473684210526
2	0.89473684210526	0.10526315789473	0.89473684210526
3	0.89473684210526	0.10526315789473	0.89473684210526
4	0.89473684210526	0.10526315789473	0.89473684210526
5	0.89473684210526	0.10526315789473	0.89473684210526

Red circles highlight the 'Exported Data' property in the 'Train' section, the 'Browse...' button, and the 'Predicted: Creditability=bad' column header in the 'EMWS1.Tree_TRAIN' window.

Predicted Probabilities come from the decision tree as expected

Oversampled Data *with* Priors

The screenshot displays the Enterprise Miner - Priors and Decisions interface. The left pane shows the project structure with 'Data Sources', 'Diagrams', 'priors', and 'Model Packages'. The main area shows the 'Exported Data - Priors' dialog box, which contains a table of data sources and their roles. The 'Train' tab is selected, showing the 'EMWS1.Tree3_TRAIN' table. The 'Exported Data' property is highlighted in the left pane. The 'EMWS1.Tree3_TRAIN' table is also highlighted in the main area, showing predicted probabilities and unadjusted probabilities for creditability.

Port	Table	Role	Data Exists
TRAIN	EMWS1.Tree3_TRAIN	Train	Yes
VALIDATE	EMWS1.Tree3_VALIDATE	Validate	No
TEST	EMWS1.Tree3_TEST	Test	No
SCORE	EMWS1.Tree3_SCORE	Score	No
TRANSACTION	EMWS1.Tree3_TRANSACTION	Transaction	No
TREE	EMWS1.Tree3_EMTREE	Tree	Yes

Property	Value
Node ID	Tree3
Imported Data	...
Exported Data	...
Train	
Variables	...
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

	Predicted: Creditability=ba	Unadjusted P: Creditability=good	Unadjusted P: Creditability=bad	Re
1	0.5107296137339	0.10526315789473	0.89473684210526	-0.10
2	0.5107296137339	0.10526315789473	0.89473684210526	-0.10
3	0.5107296137339	0.10526315789473	0.89473684210526	-0.10
4	0.5107296137339	0.10526315789473	0.89473684210526	-0.10

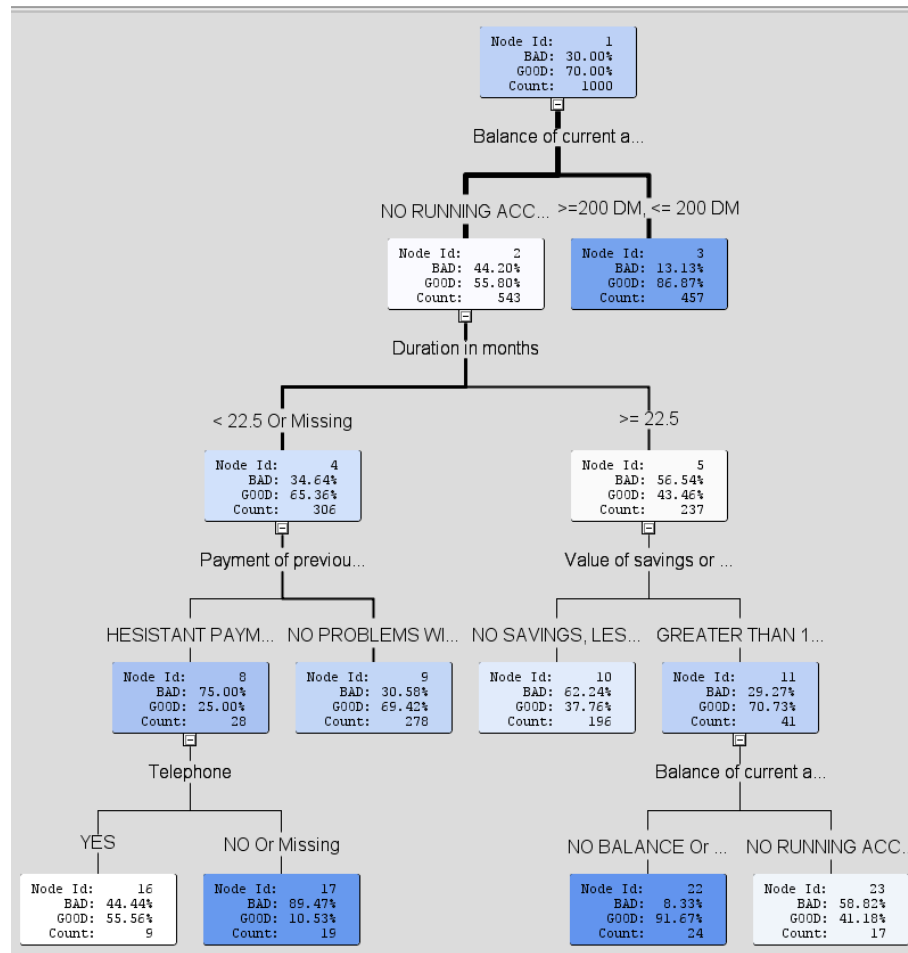
Predicted Probabilities are adjusted according to the priors. The tree is pruned according to those adjustments too. Default cutoff probability is still 0.5!

Oversampled Data with Priors

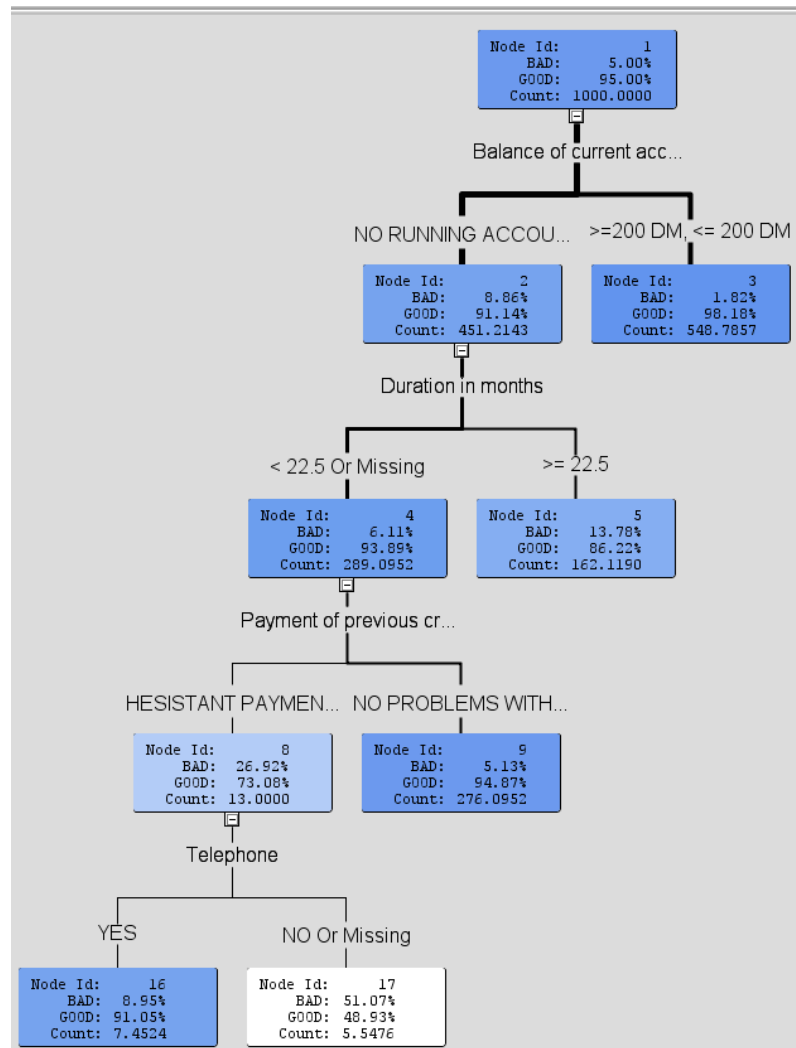
Predicted: Creditability=good	Predicted: Creditability=bad	Unadjusted P: Creditability=good	Unadjusted P: Creditability=bad	Residual: Creditability=good	Residual: Creditability=bad	Decision
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	0.89473684210527	-0.89473684210526	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	0.89473684210527	-0.89473684210526	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.48927038626609	0.5107296137339	0.10526315789473	0.89473684210526	-0.10526315789473	0.10526315789473994	BAD
0.86224115141724	0.13775884858275	0.43459915611814	0.56540084388185	0.5654008438818601	-0.56540084388185	GOOD
0.86224115141724	0.13775884858275	0.43459915611814	0.56540084388185	0.5654008438818601	-0.56540084388185	GOOD
0.86224115141724	0.13775884858275	0.43459915611814	0.56540084388185	-0.43459915611814	0.43459915611815003	GOOD
0.86224115141724	0.13775884858275	0.43459915611814	0.56540084388185	0.5654008438818601	-0.56540084388185	GOOD

Default cutoff probability is still 0.5!

Oversampled Data with No Priors



Oversampled Data *with* Priors



Using Inv. Priors as Decision Weights

- To emphasize rare events in a modelling context, we may want to increase the “profit” of making a correct prediction of the rare event.
- The easiest way to do this is to weight the decisions with a profit (or cost – make errors negative) matrix:
- Priors: RareEvent = 0.02, CommonEvent = 0.98

Decision Weights to emphasize correct classification of rare events		Predicted	
		RareEvent	CommonEvent
Actual	RareEvent	$1/0.02 = \mathbf{50}$	0
	CommonEvent	0	$1/0.98 = \mathbf{1.02}$

Creating Inv. Prior Decision Weights

Property Value

General

Node ID Ids3

Imported Data ...

Exported Data ...

Notes ...

Train

Output Type View

Role Raw

Rerun No

Summarize No

Drop Map Variables Yes

Columns

Variables ...

Decisions ...

Refresh Metadata ...

Advisor Basic

Advanced Options ...

Data

Data Selection Data Source

Sample Default

Sample Options ...

Data Source

Data Source CREDIBILITY ...

General

General Properties

Decision Processing - CREDIBILITY_priors

Target **Prior Probabilities** Decisions Decision Weights

Do you want to enter new prior probabilities?

☒ Yes ☐ No

Level	Count	Prior	Adjusted Prior
GOOD	700	0.7	0.95
BAD	300	0.3	0.05

If first time, click "Build"

Creating Inv. Prior Decision Weights

Decision Processing - CREDIBILITY_priors_invDecisionWeights

Targets Prior Probabilities **Decisions** Decision Weights

Do you want to use the decisions:

☒ Yes ☐ No

Default with Inverse Prior Weights

Decision Name	Label	Cost Variable	Constant
DECISION1	GOOD	< None >	0.0
DECISION2	BAD	< None >	0.0

Add
Delete
Delete All
Reset
Default

OK Cancel

Creating Inv. Prior Decision Weights

Decision Processing - CREDIBILITY_priors_invDecisionWeights

Targets Prior Probabilities Decisions **Decision Weights**

Select a decision function:

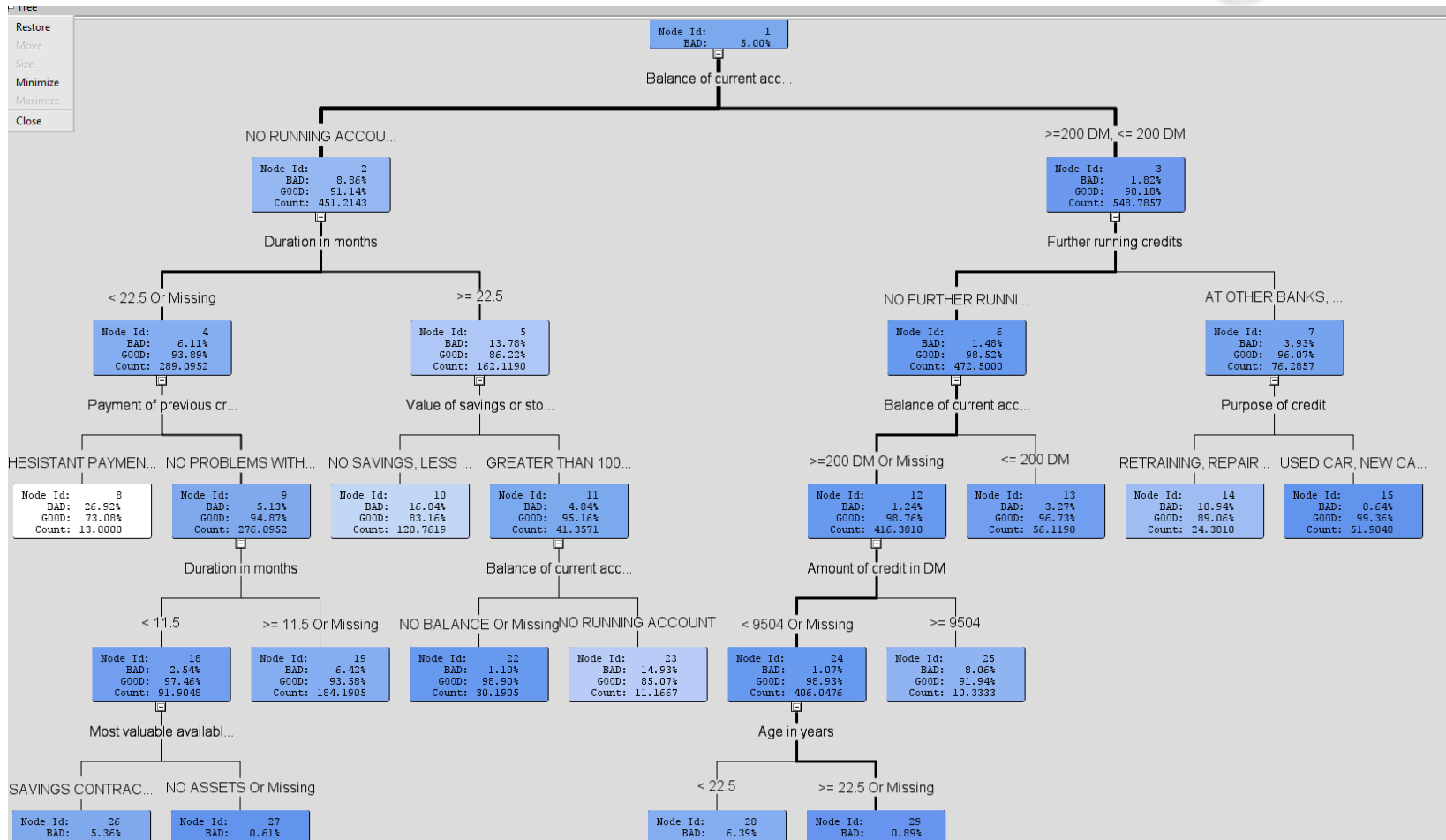
☒ Maximize ☐ Minimize

Enter weight values for the decisions.

Level		DECISION1	DECISION2
GOOD	...	1.05263157...	0.0
BAD	...	0.0	20.0

OK Cancel

Oversampled Data with Priors and Decision Weights



Oversampled Data with Priors

Predicted: Creditability=bad ▾	Unadjusted P: Creditability=good	Unadjusted P: Creditability=bad	Residual: Creditability=good	Residual: Creditability=bad	Decision: Creditability
0.26923076923076	0.25	0.75	0.75	-0.75	BAD
0.26923076923076	0.25	0.75	-0.25	0.25	BAD
0.26923076923076	0.25	0.75	-0.25	0.25	BAD
0.26923076923076	0.25	0.75	-0.25	0.25	BAD
0.26923076923076	0.25	0.75	-0.25	0.25	BAD
0.26923076923076	0.25	0.75	0.75	-0.75	BAD
0.26923076923076	0.25	0.75	0.75	-0.75	BAD

Cutoff probability is now the population probability of the rare event, here $p=0.05$!