Ask the Expert

Model Selection Techniques in SAS® Enterprise Guide® and SAS® Enterprise Miner™

Goals

- Increase awareness of and comfort with capabilities in SAS® for doing model selection in
  - SAS® Enterprise Guide®
  - SAS® Enterprise Miner™
- Share resources for learning more

AGENDA

Model Selection Techniques

- What is model selection?
- Why is it important?
- How to do model selection using
  - SAS® Enterprise Guide®
  - SAS® Enterprise Miner™

What?

Model Selection

The task of selecting a statistical model from a set of candidate models in order to meet our objectives.

Model Selection and Variable Selection

What?

Model Selection GOALS

When we have many predictors it can be difficult to find a good model.

- Which main effects do we include?
- Which interactions do we include?
- Which modeling algorithm do we use?

Model selection tries to “simplify” this task.

Why?

Model Selection

Essentially, all models are wrong, but some are useful.

George E. Box

Why?

Opposing Goals

Model Selection

- Good fit, good in-sample prediction
  - Include many variables
- Parsimony:
  - Keep cost of data collection low, interpretation simple

[Model Selection in Regression, Michael Friendly]

How?

It Depends!

What is the situation?

Comparing models using the same algorithm

OR

Comparing models from different algorithms

How?


Mallows C
ASE
Gauss
Profit Loss
Lift Charts
KS Statistic
R-Squared
Captured Response
SB
BIC
AIC
Gini Coefficient
Misclassification
Backward
Kullback-Leibler
Pearson
RMSE
Adjusted R-Square
MSE
Stepwise
Forward
PRESS
Gains Charts

How?
Model Selection Criteria

- Goodness-of-Fit Statistics
  - R^2
  - Adjusted R^2
  - MSE or RMSE
  - Mallows C, PRESS *(Linear Regression)*
- Information Criteria
  - AIC
  - SC (SBC or BIC)
- Assessment/Visualizations
  - Misclassification *(logistic, classification trees)*
  - ROC
  - Lift
  - Gains Charts

Criterion

Goodness-of-Fit Statistics

R^2

- R Square is a measure of model accuracy

(bad / low) 0 < R Square < 1 (good / high)

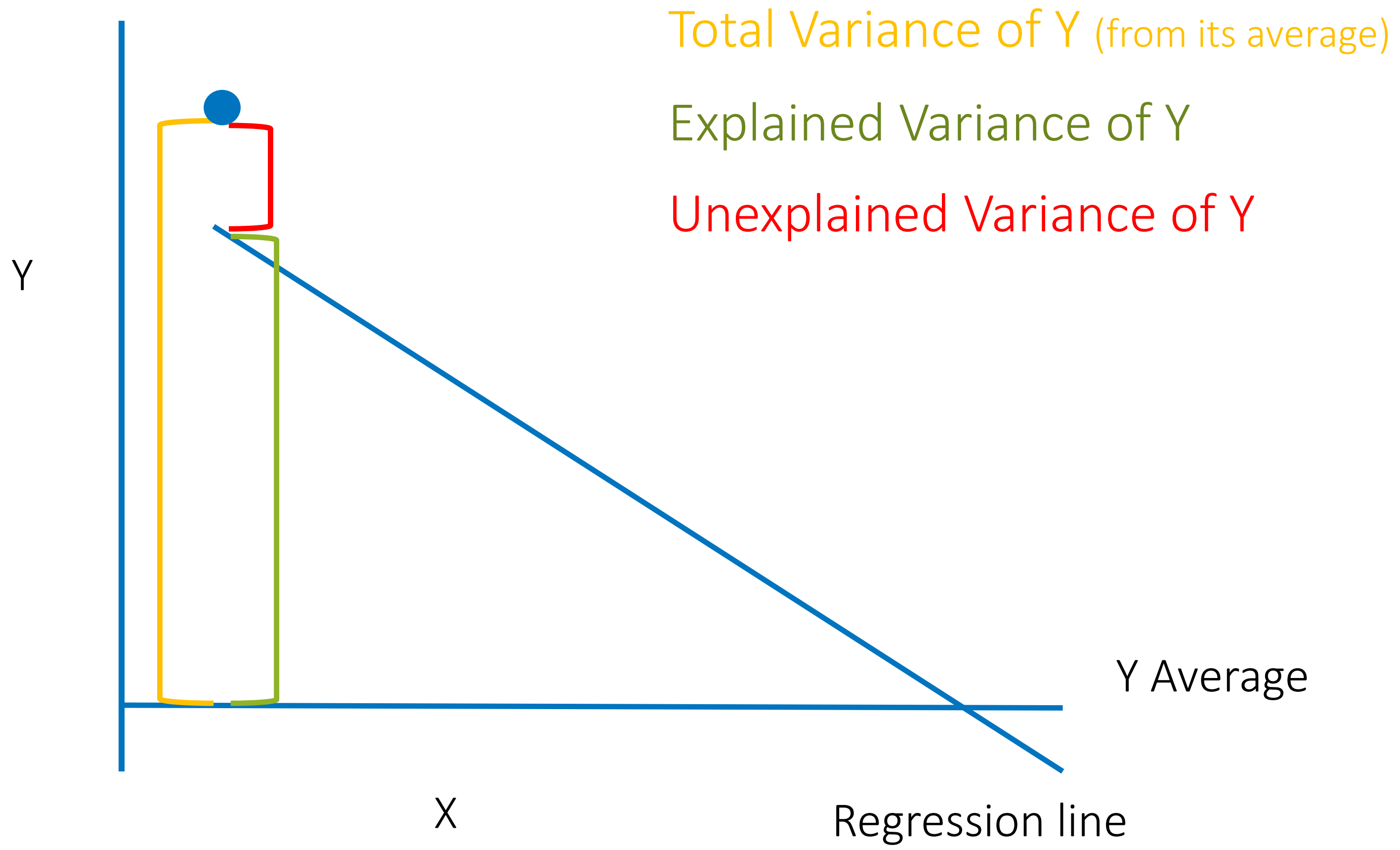
- It is a relative measure. Is 0.5078 good or bad?
    - What are you comparing it to?

Linear Regression | R Square Explained

R Square =

$$\frac{\text{Exp. Variance Y}}{\text{Total Variance Y}}$$

- Sum for all the data points
- The proportion of variability in the dependent variable explained by the independent variables.
- Always increases with the addition of new variables



Criterion

Goodness-of-Fit Statistics

Adjusted R^2

- Adjusted R square = R square (penalized for number of predictors)
- Helps you to remove needless complexity
- Adjusted R^2 than can go up or down depending on whether the addition of another variable adds or does not add to the explanatory power of the model.
- Adjusted R^2 will always be lower than unadjusted.

[ONE MORE TIME ABOUT R2 MEASURES OF FIT]

Information Criterion

Akaike Information Criterion (AIC)

- Measures the difference between a given model and the “true” underlying model
- Measure of relative quality of the model
- Trade-off between goodness of fit and complexity of the model
- Smaller is better

Criterion

Information Criterion

Schwarz Criterion (SBC)

- Based on likelihood function
- Closely related to AIC
- Penalty term is larger in SBC than AIC
- Generally speaking AIC will pick a larger model than SBC
- Also know as **Bayesian Information Criterion** (**BIC**)
- Smaller is better

Criterion

Assessment / Visualizations

Misclassification

The misclassification rate calculates the proportion of an observation being allocated to the incorrect group

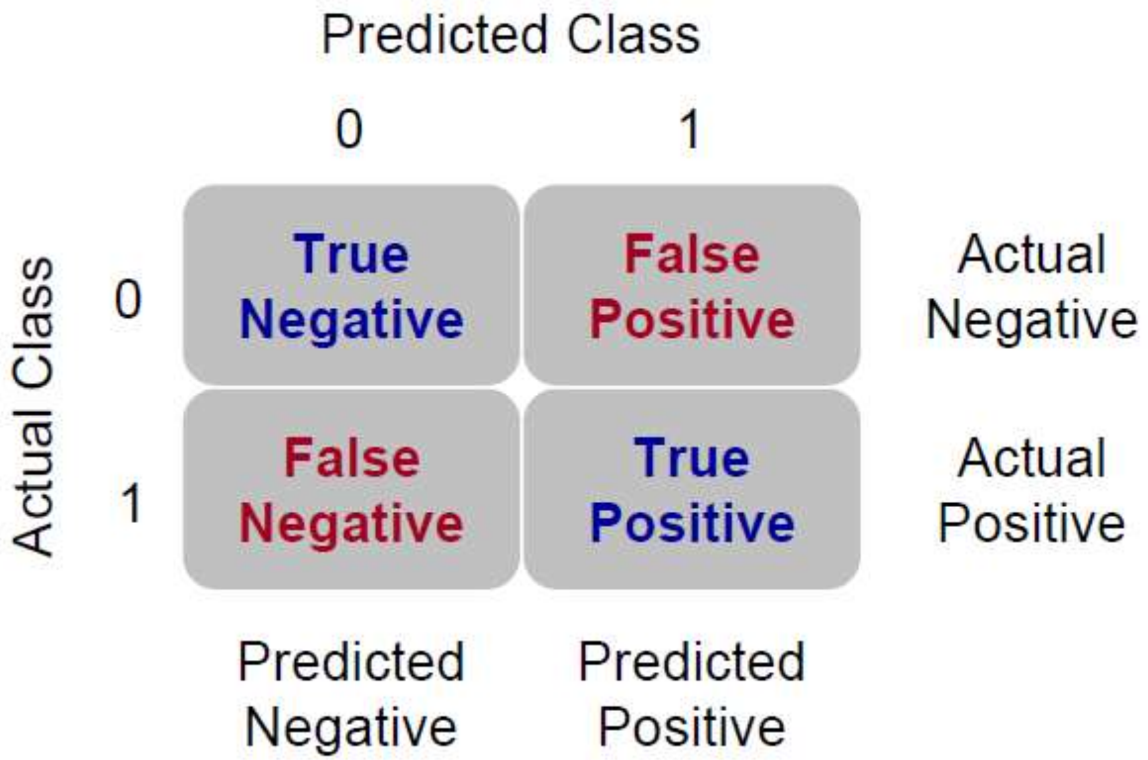
$$\frac{\#\ of\ Incorrect}{Total}$$

Criterion

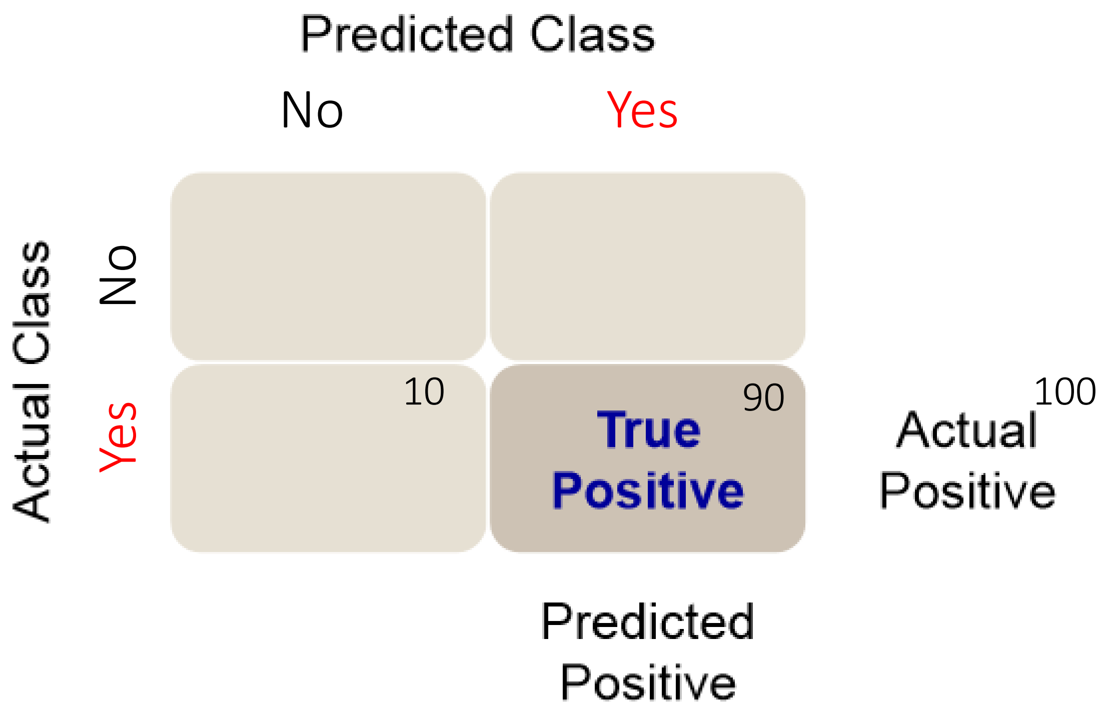
Assessment/
Visualizations

Confusion Matrix

Extension of Misclassification Rate



Confusion matrix | **Sensitivity**



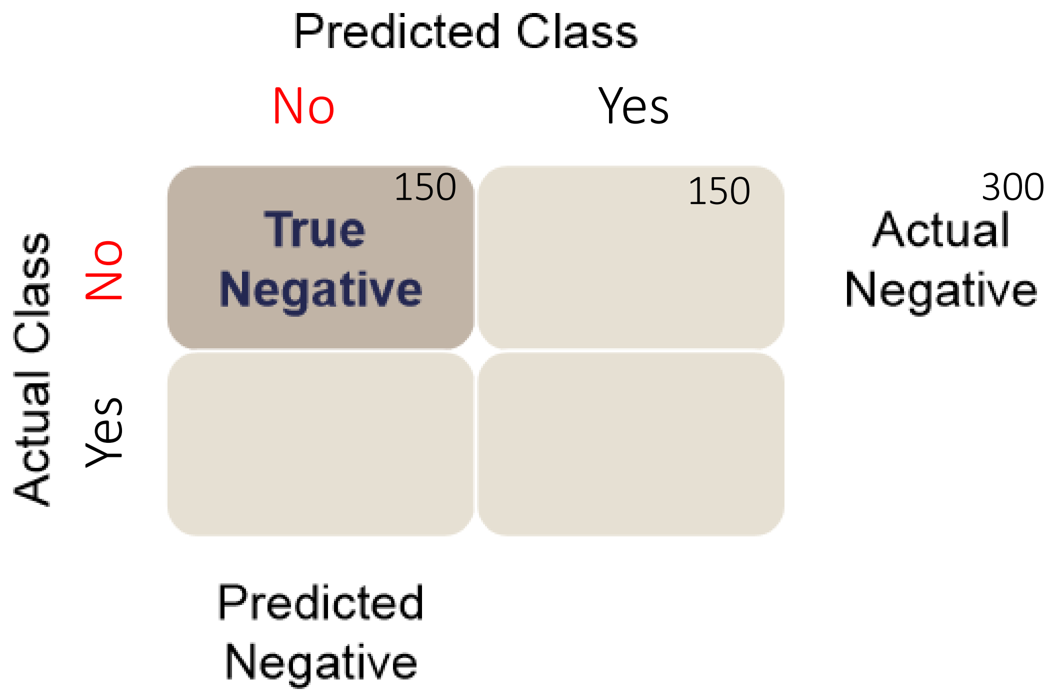
$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{Total Actual Positives}}$$

$$= \frac{\text{\# of Accurately Predicted Yes}}{\text{\# of Actual Yes}}$$

Example: 100 Yes, 90 predicted correctly

90/100 = .90 sensitivity

Confusion matrix | Specificity



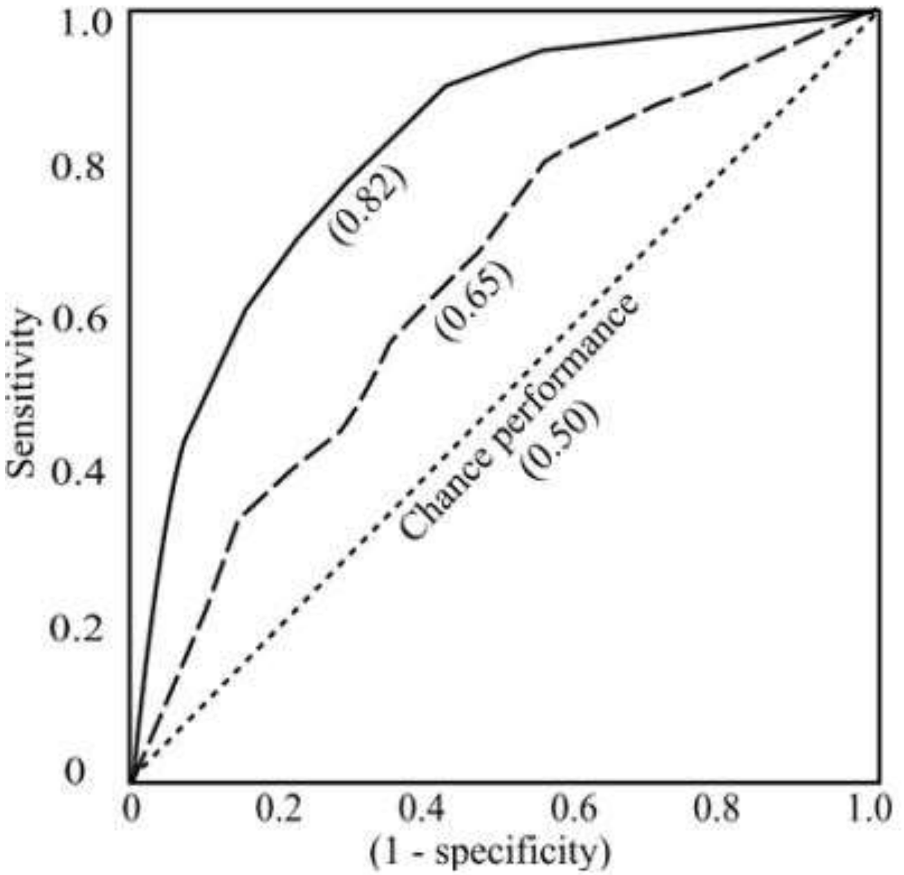
$$\text{Specificity} = \frac{\text{True Negatives}}{\text{Total Actual Negatives}}$$

$$= \frac{\text{\# of Accurately Predicted No}}{\text{\# of Actual No}}$$

Example: 300 No, 150 predicted correctly

150/300 = .50 specificity

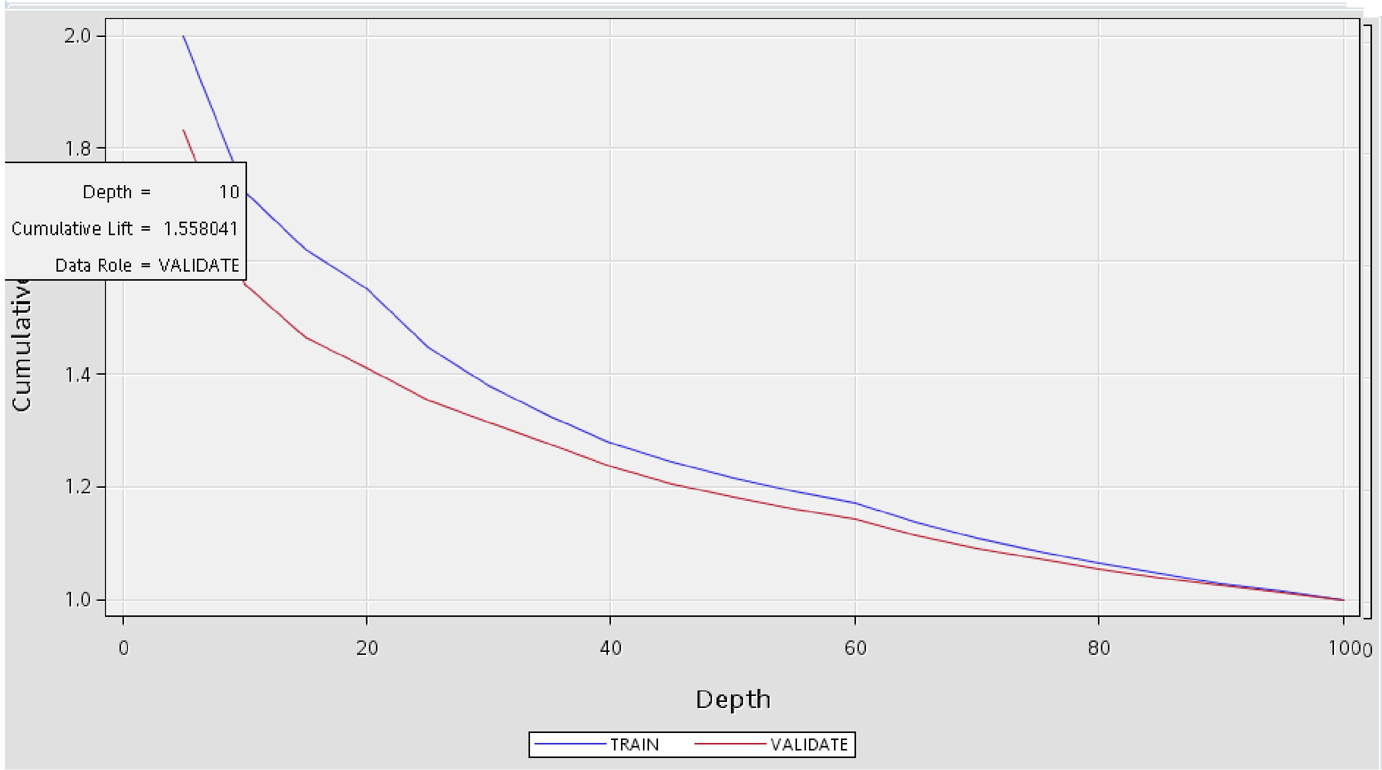
ROC Curves – (Receiver Operating Characteristic)



- Reflect tradeoff between sensitivity and specificity
- A model with high predictive accuracy will rise quickly (moving from left to right) indicating that higher levels sensitivity can be achieved without sacrificing much specificity
- (.82) = area under the curve

Assessment
Lift Charts

A lift chart is a method of visualizing the improvement that you get from using a data mining model, when you compare it to random guessing.



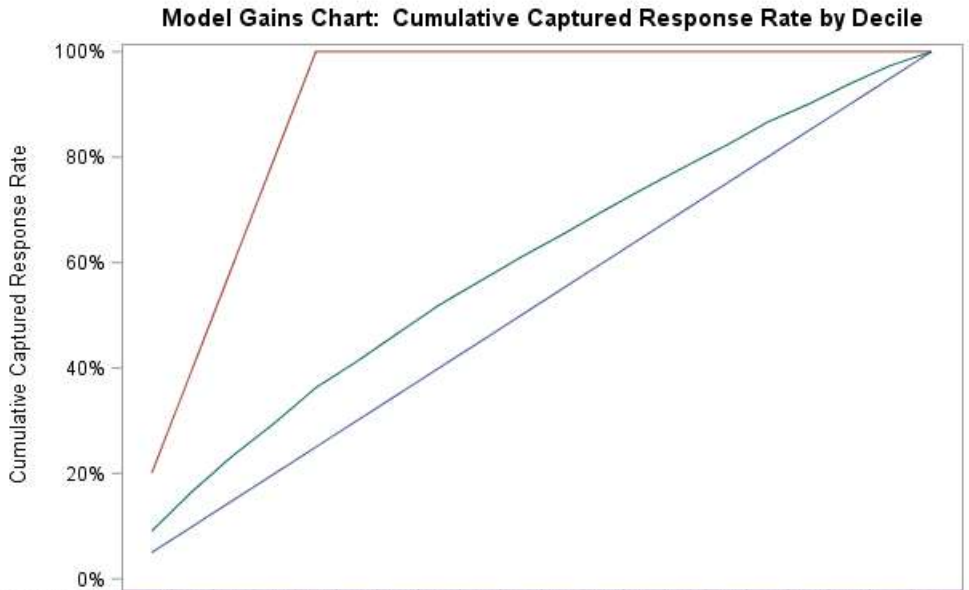
Example

- 25% have donated
- Build a model to predict which people will donate for an upcoming postcard campaign
- Rank all from most to least likely to donate based on model
- Compare actual % of donated in each percentile(x-axis) to 25% baseline for entire data set

Lift = (Expected Response In A Specific Lot Prospects Using Predictive Model) / (Expected Response In A Random Lot Prospects Without Using Predictive Model)

Assessment
Gains Charts

- The cumulative gains chart shows the percentage of the overall number of cases in a given category "gained" by targeting a percentage of the total number of cases.
- How many more responses do we gain by using our model?



Cumulative Lift	1.811	1.667	1.568	1.479	1.45	1.375	1.332	1.298	1.255	1.221	1.187	1.163	1.139	1.117	1.096	1.083	1.059	1.043	1.025	1
Lift	1.811	1.523	1.368	1.214	1.335	0.997	1.073	1.065	0.907	0.915	0.844	0.903	0.857	0.83	0.804	0.879	0.68	0.771	0.699	0.523
Model	9%	17%	24%	30%	36%	41%	47%	52%	56%	61%	65%	70%	74%	78%	82%	87%	90%	94%	97%	100%
Best	20%	40%	60%	80%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Percentile	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%

Honest Assessment
Data Splitting

Evaluating the effectiveness of a model based on a hold out sample.

Steps:

1. Split data into Training and Validation data sets
2. Create a model using the Training dataset
3. Assess model accuracy using Validation dataset & your favorite criteria

Data

Donor_Raw_Data

People likely to donate to a charity

- Y=TARGET_B
- N = 19,372
- Variables = 50 (47 possible inputs)

Alphabetic List of Varia				
#	Variable	Type	Len	Form
37	CARD_PROM_12	Num	8	
8	CLUSTER_CODE	Char	2	
3	CONTROL_NUMBER	Char	8	
10	DONOR_GENDER	Char	3	
41	FILE_AVG_GIFT	Num	8	
42	FILE_CARD_GIFT	Num	8	
21	FREQUENCY_STATUS_97NK	Num	8	
9	HOME_OWNER	Char	3	
48	IM_DONOR_AGE	Num	8	2.
46	IM_INCOME_GROUP	Num	8	2.
50	IM_MONTHS_SINCE_LAST_PROM_RESP	Num	8	2.
44	IM_WEALTH_RATING	Num	8	2.
5	IN_HOUSE	Num	8	
36	LAST_GIFT_AMT	Num	8	
32	LIFETIME_AVG_GIFT_AMT	Num	8	
28	LIFETIME_CARD_PROM	Num	8	
30	LIFETIME_GIFT_AMOUNT	Num	8	
31	LIFETIME_GIFT_COUNT	Num	8	
33	LIFETIME_GIFT_RANGE	Num	8	
34	LIFETIME_MAX_GIFT_AMT	Num	8	
35	LIFETIME_MIN_GIFT_AMT	Num	8	
29	LIFETIME_PROM	Num	8	
14	MEDIAN_HOME_VALUE	Num	8	
15	MEDIAN_HOUSEHOLD_INCOME	Num	8	
40	MONTHS_SINCE_FIRST_GIFT	Num	8	
39	MONTHS_SINCE_LAST_GIFT	Num	8	
4	MONTHS_SINCE_ORIGIN	Num	8	
13	MOR_HIT_RATE	Num	8	
47	M_DONOR_AGE	Num	8	
45	M_INCOME_GROUP	Num	8	
49	M_MONTHS_SINCE_LAST_PROM_RESP	Num	8	
43	M_WEALTH_RATING	Num	8	
38	NUMBER_PROM_12	Num	8	
12	OVERLAY_SOURCE	Char	1	
16	PCT_OWNER_OCCUPIED	Num	8	
18	PEP_STAR	Num	8	
17	PER_CAPITA_INCOME	Num	8	
11	PUBLISHED_PHONE	Num	8	
20	RECENCY_STATUS_96NK	Char	5	
25	RECENT_AVG_CARD_GIFT_AMT	Num	8	
23	RECENT_AVG_GIFT_AMT	Num	8	
27	RECENT_CARD_RESPONSE_COUNT	Num	8	
24	RECENT_CARD_RESPONSE_PROP	Num	8	
26	RECENT_RESPONSE_COUNT	Num	8	
22	RECENT_RESPONSE_PROP	Num	8	
19	RECENT_STAR_STATUS	Num	8	
7	SES	Char	4	
1	TARGET_B	Num	8	
2	TARGET_D	Num	8	
6	URBANICITY	Char	4	

DATA

First Things First

Impute missing values

Categorical

```
if donor_gender in ('U','A') then
    donor_gender='U';

if SES='?' then
    SES='5';

if URBANICITY='?' then
    URBANICITY='M';
```

Set Gender to Unknown, SES to Level 5 (Unknown), Urbanity to M (missing)

Continuous

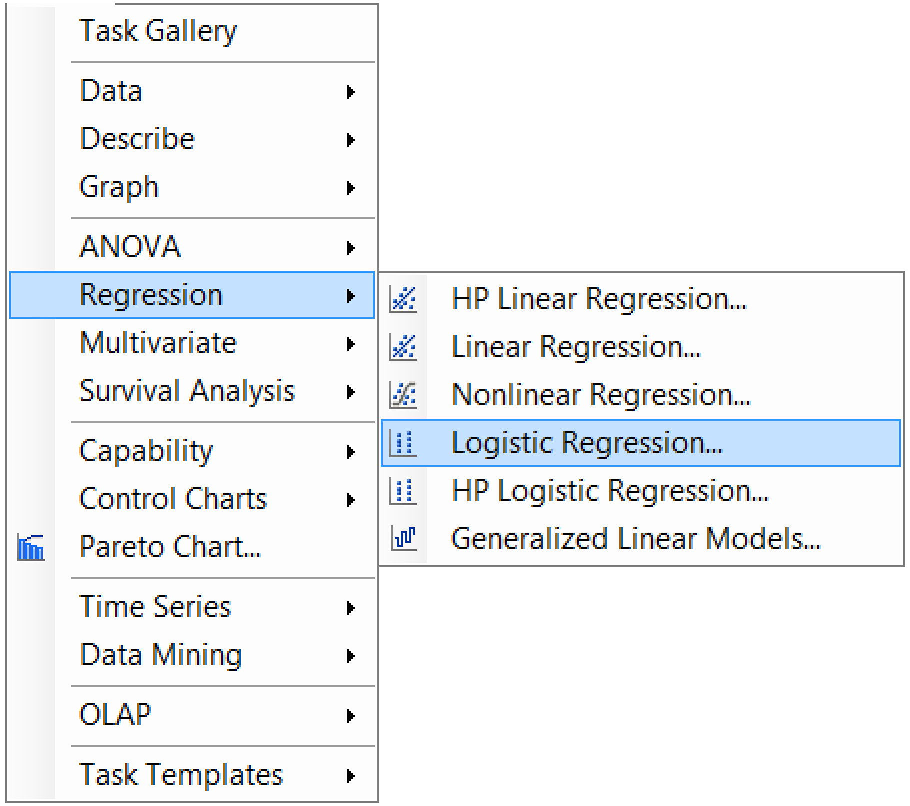
```
proc HPIMPUTE data=Donor.donor_raw_data out=out1;
     input  wealth_rating income_group donor_age months_since_last_prom_resp;
     impute wealth_rating / method=random;
     impute income_group / method=random;
     impute donor_age /method=random;
     impute months_since_last_prom_resp / method=random;
run;
```

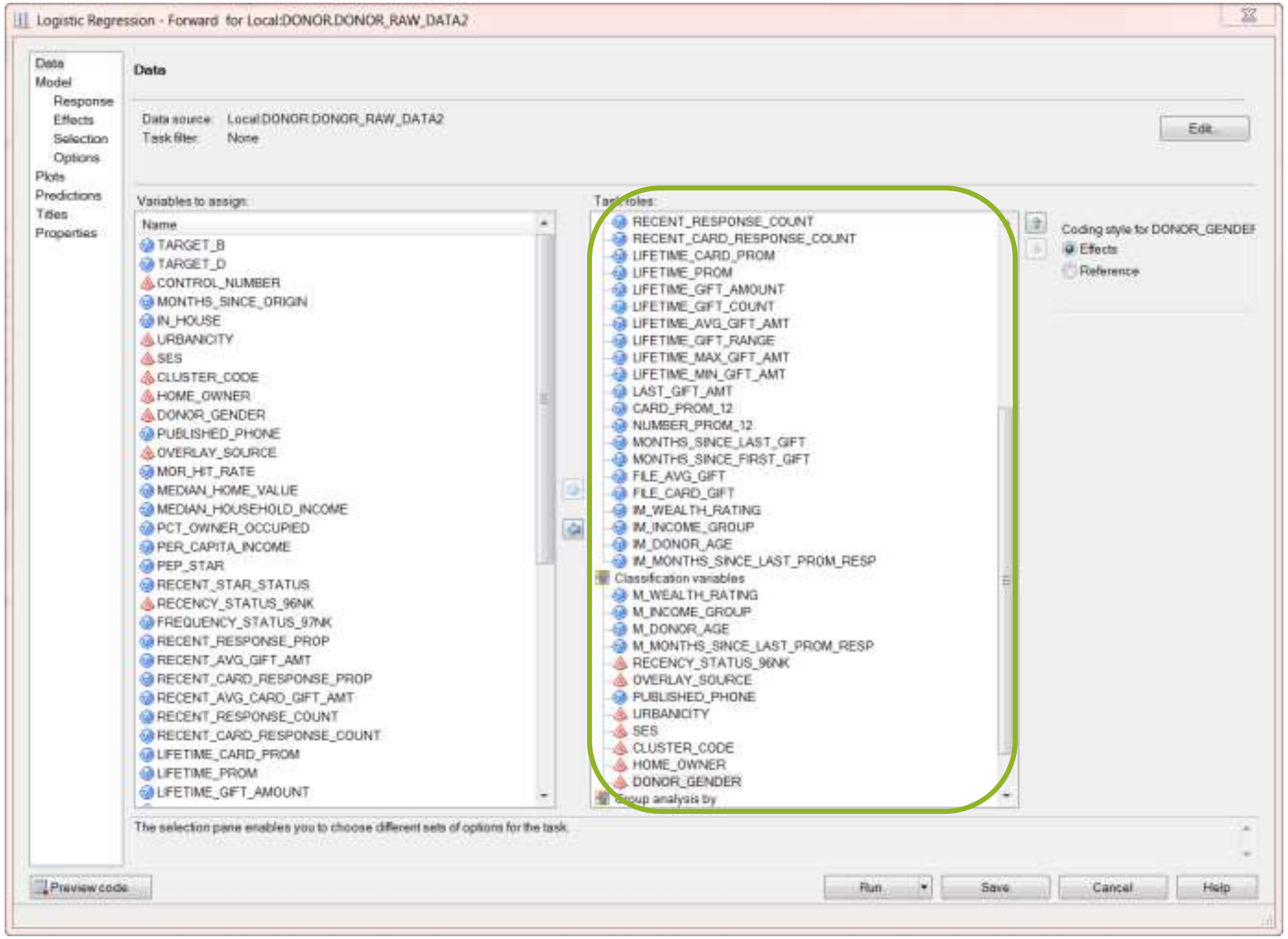
MEAN, RANDOM, PMEDIAN or Constant Value

SAS® Enterprise Guide®

SAS® Enterprise Guide®

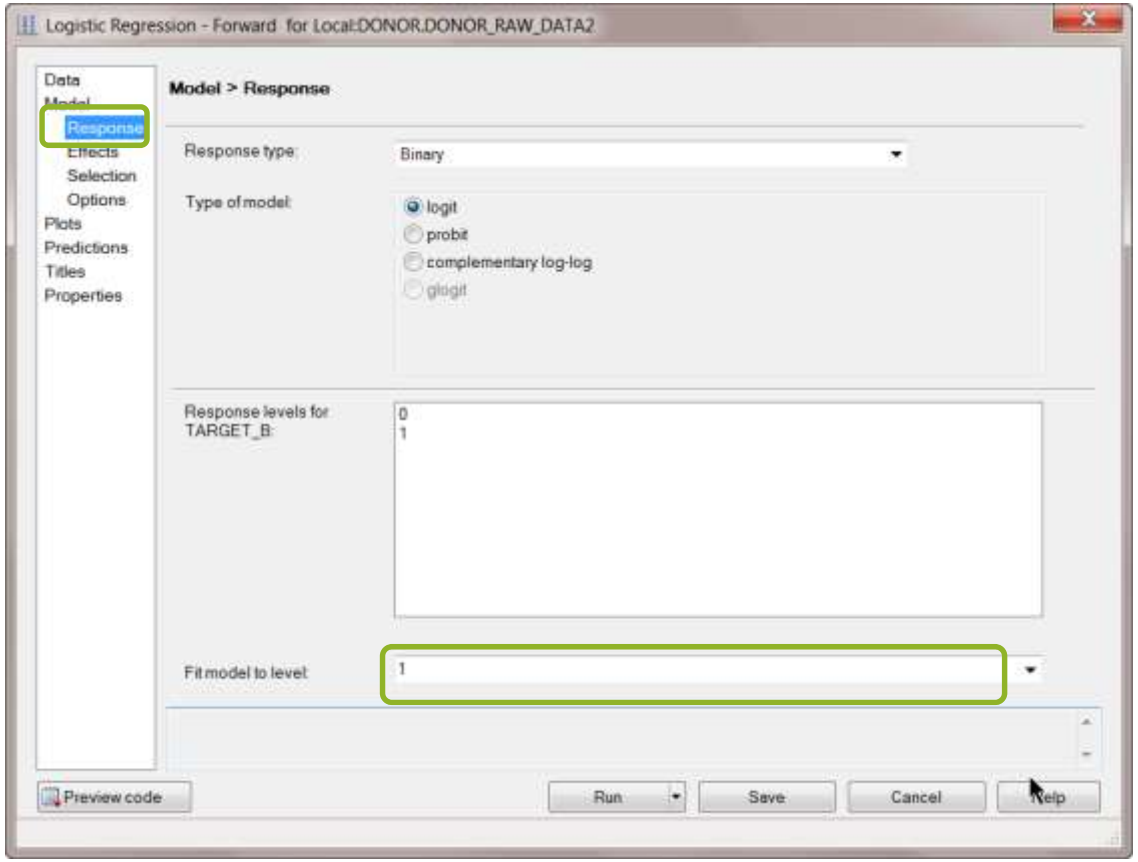
Logistic Regression

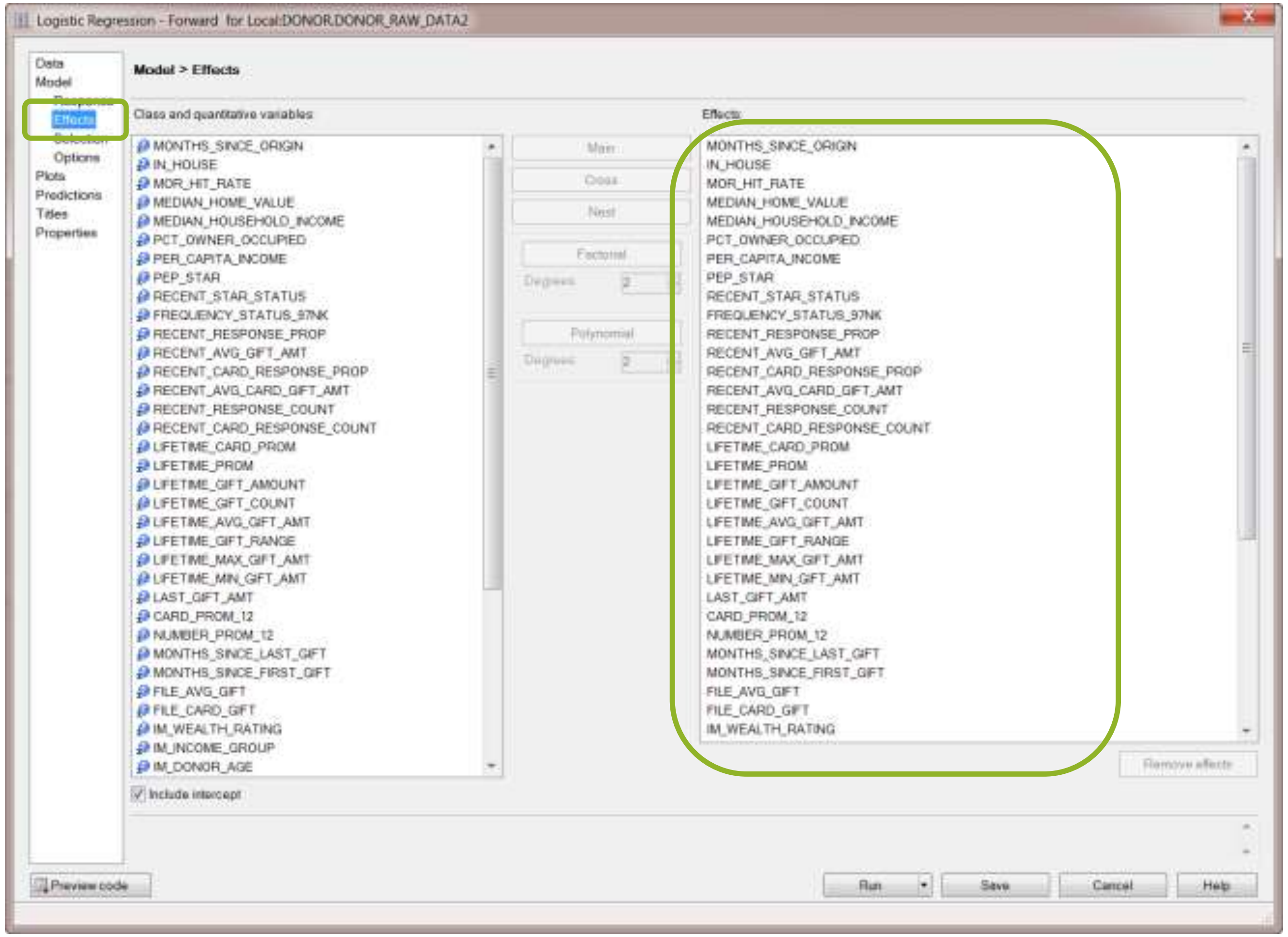




SAS® Enterprise Guide®

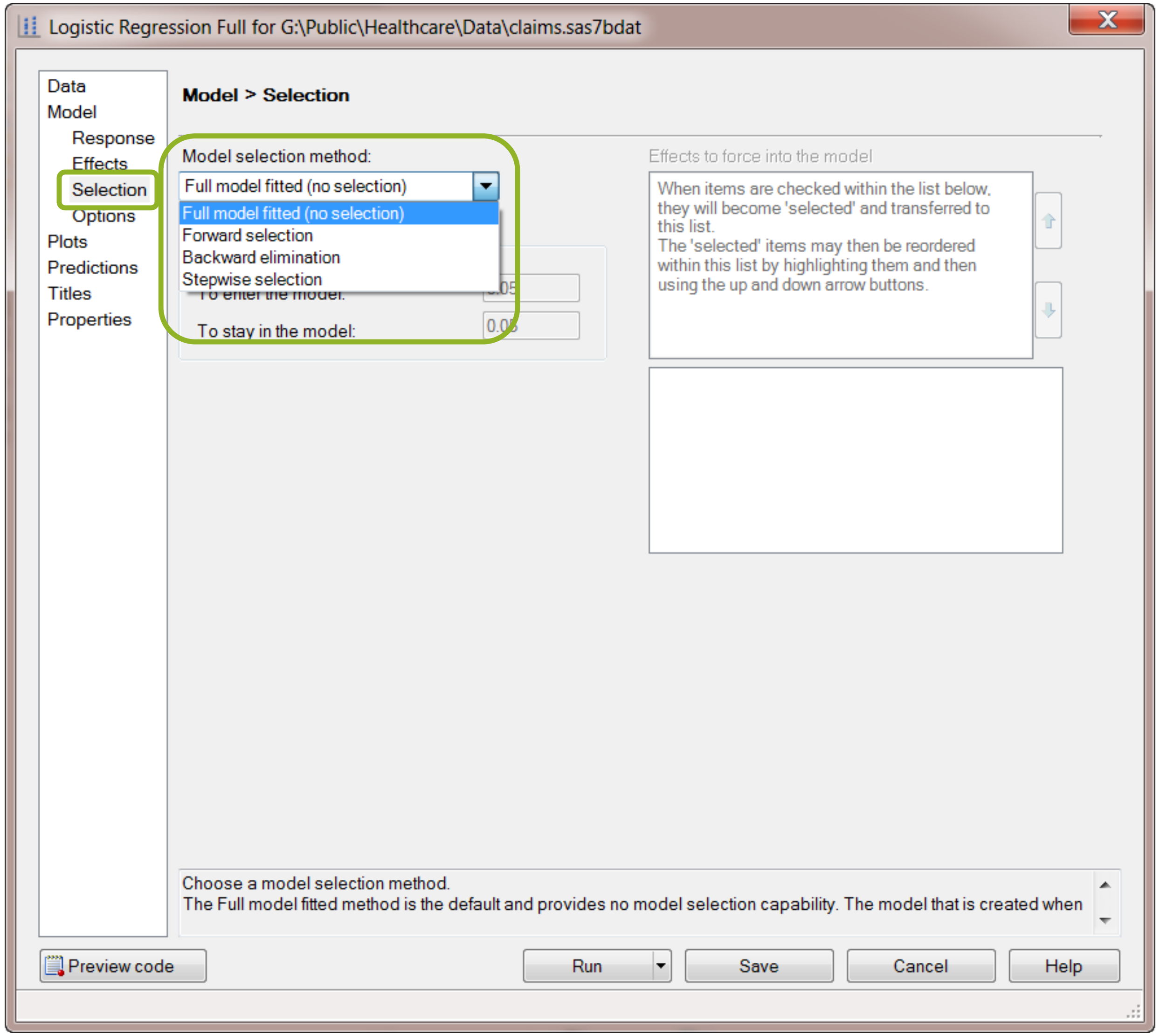
Logistic Regression





SAS® Enterprise Guide®

Logistic Regression



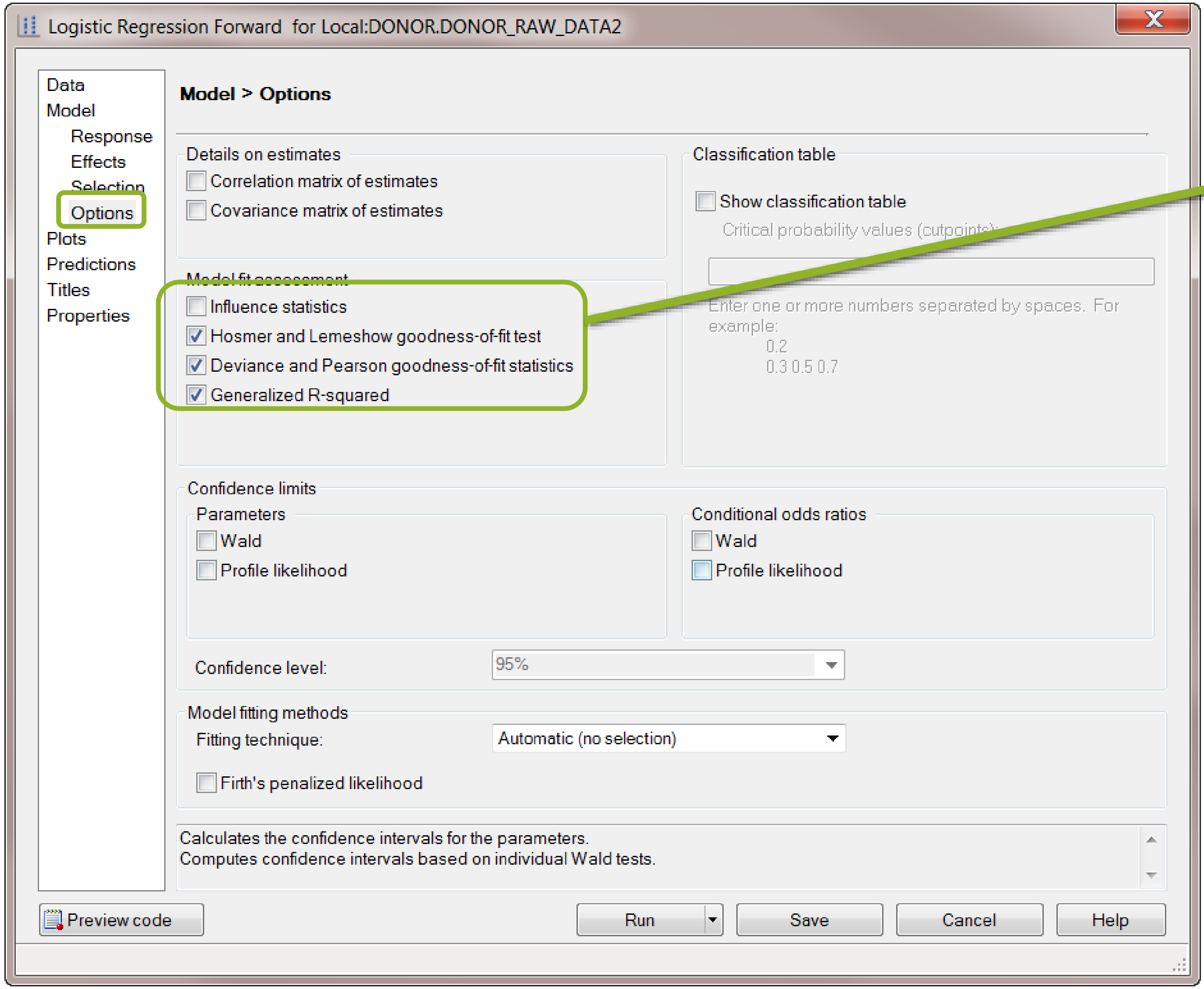
- Selection=None
- Selection=Forward
- Selection=Backward
- Selection=Stepwise
- Selection=Score

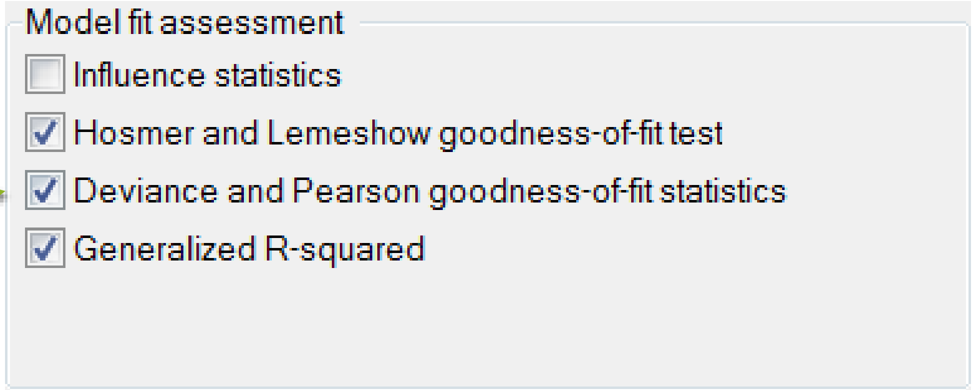
❖ Ones in BLUE available in SAS Enterprise Guide

[Variable Selection Methods in Proc Logistic Documentation]

SAS® Enterprise Guide®

Logistic Regression - Criterion





AIC and SBC are automatically part of the output (SBC is labeled as SC)

[Model Fitting in Proc Logistic Documentation]

SAS® Enterprise Guide®

Logistic Regression

Variable Name	Stepwise	Backward	Forward
CLUSTER_CODE		*	
FREQUENCY_STATUS_97N	*	*	*
HOME_OWNER	*	*	*
IM_WEALTH_RATING	*	*	*
IN_HOUSE		*	
LIFETIME_CARD_PROM	*		*
M_WEALTH_RATING	*	*	*
MEDIAN_HOME_VALUE	*	*	*
MONTHS_SINCE_FIRST_GIFT	*	*	*
MONTHS_SINCE_LAST_GIFT	*	*	*
NUMBER_PROM_12		*	
PEP_STAR	*	*	*
RECENT_AVG_GIFT_AMT	*	*	*
RECENT_CARD_RESPONSE_COUNT			*
RECENT_CARD_RESPONSE_PROP	*	*	*
SES	*		*
Number of Variables	12	13	13

SAS® Enterprise Guide®

Logistic Regression

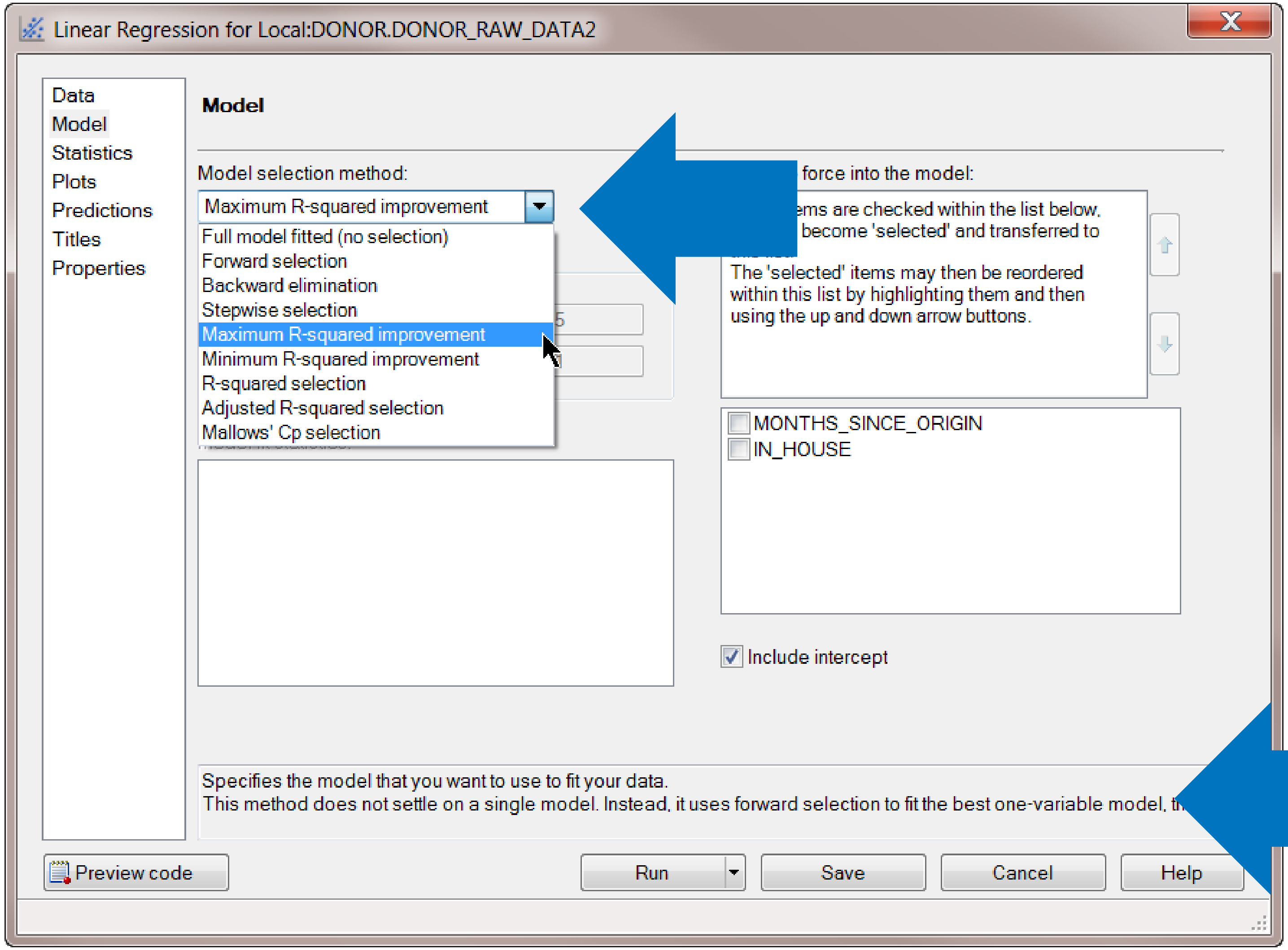
	# Terms	R^2	Max Rescaled R^2	AIC	SBC
Model					
Full	38	0.0357	0.0528	21183	21576
Stepwise	12	0.0344	0.0503	21150	21316
Backward	13	0.0341	0.0505	21155	21313
Forward	13	0.0344	0.0510	21150	21315
*2-way Interactions	30	0.0424	0.0628	21021	21312
*2 degree Polynomial Terms	18	0.0365	0.0540	21111	21284
*2-Way + 2 degree Poly	31	0.2622	0.3497	21039	21339
		Largest	Largest	Smallest	Smallest

PROC Logistic calculates a Pseudo R^2 and Max Rescaled R^2 which divides the Pseudo by the upper bound.

* Using Forward Selection Method

SAS® Enterprise Guide®

Variable Selection Methods in SAS/Stat Proc REG



- Forward
- Backward
- Stepwise
- LASSO
- LARS
- MAXR
- MINR
- RSQUARE
- CP
- ADJRSQ
- SCORE

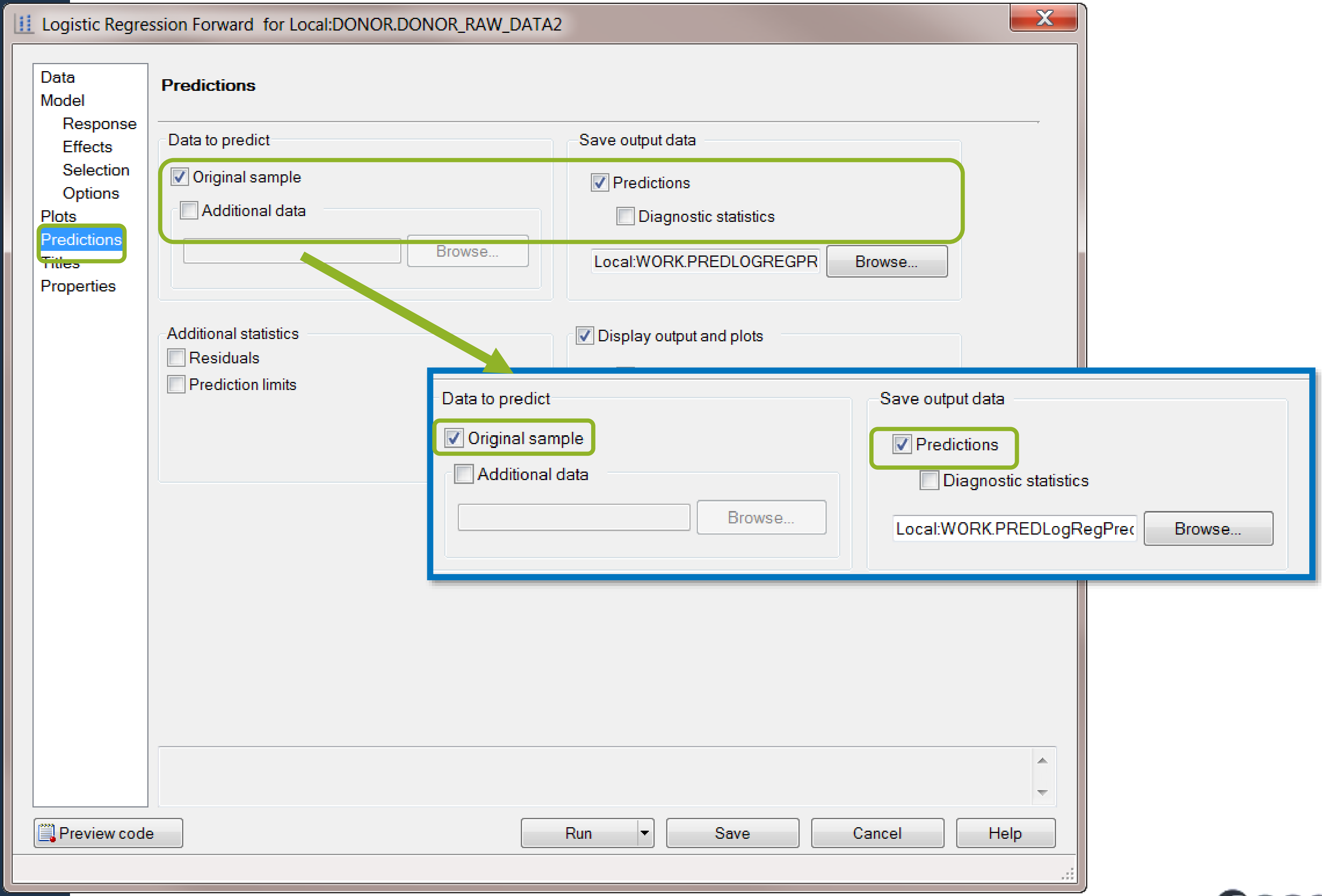
❖ Ones in BLUE available in SAS Enterprise Guide

Only for numeric variables

SAS® Enterprise Guide®

Misclassification

Method 1



SAS® Enterprise Guide®

Misclassification

Method 1

Tasks→ Describe→ Table Analysis

FROM	_INTO_	IP_0	IP_1
0	0	0.7780306106	0.2219693894
1	0	0.6447591583	0.3552408417
0	0	0.5527546616	0.4472453384
0	0	0.5759559523	0.4240440477
0	0	0.8314731824	0.1685268176
0	0	0.7520974014	0.2479025986
0	0	0.8080213456	0.1919786544
1	0	0.5767807105	0.4232192895
0	0	0.688637872	0.311362128
1	0	0.5739493227	0.4260506773
0	0	0.620249476	0.379750524
0	0	0.9164326064	0.0835673936
0	0	0.6065249703	0.3934750297
0	0	0.837057543	0.162942457
1	0	0.756648627	0.243351373

Dataset created with 4 new columns

Table of _FROM_ by _INTO_			
FROM(Formatted Value of the Observed Response)	_INTO_(Formatted Value of the Predicted Response)		
	0	1	Total
0	14493	36	14529
1	4790	53	4843
Total	19283	89	19372

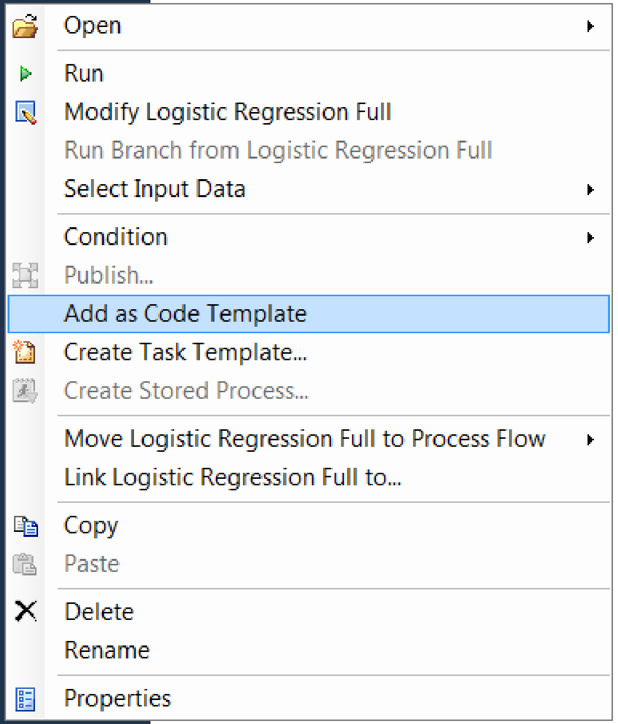
Misclassification = (4790+36)/19372 =0.2491

SAS® Enterprise Guide®

Misclassification

Method 2

Right Mouse Click on Logistic Regression Node




```
PROC LOGISTIC DATA=DONOR.DONOR_RAW_DATA2
        PLOTS(ONLY)=ALL
        OUTMODEL=DONOR.log_forw
        ;
        ...
Run;

proc logistic inmodel=DONOR.log_forw;
        score data=DONOR.DONOR_RAW_DATA2
out=score1 fitstat;
```

SAS® Enterprise Guide®

Misclassification

```
proc logistic inmodel=DONOR.log_forw;
      score data=DONOR.DONOR_RAW_DATA2 out=score1
      fitstat;
```

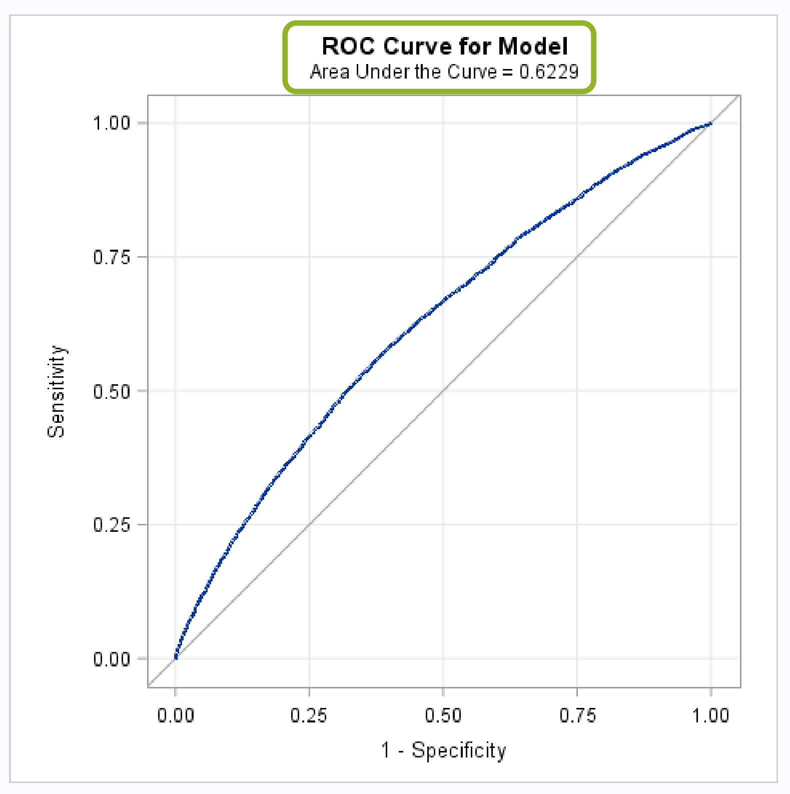
Method 2

Data Set	Total Frequency	Log Likelihood	Error Rate	
DONOR.DONOR_RAW_DATA2	19372	-10554.3	0.2491	21150

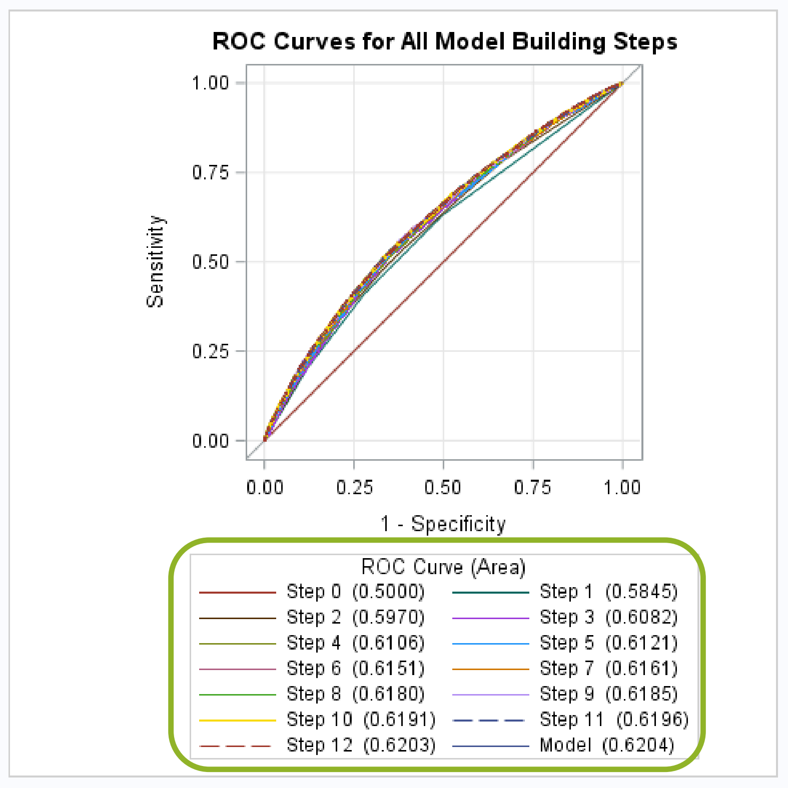
SAS® Enterprise Guide®

ROC Curves – (Receiver Operating Characteristic)

Automatically create for Logistic Regression



Full Model



Stepwise Model

SAS® Enterprise Guide®

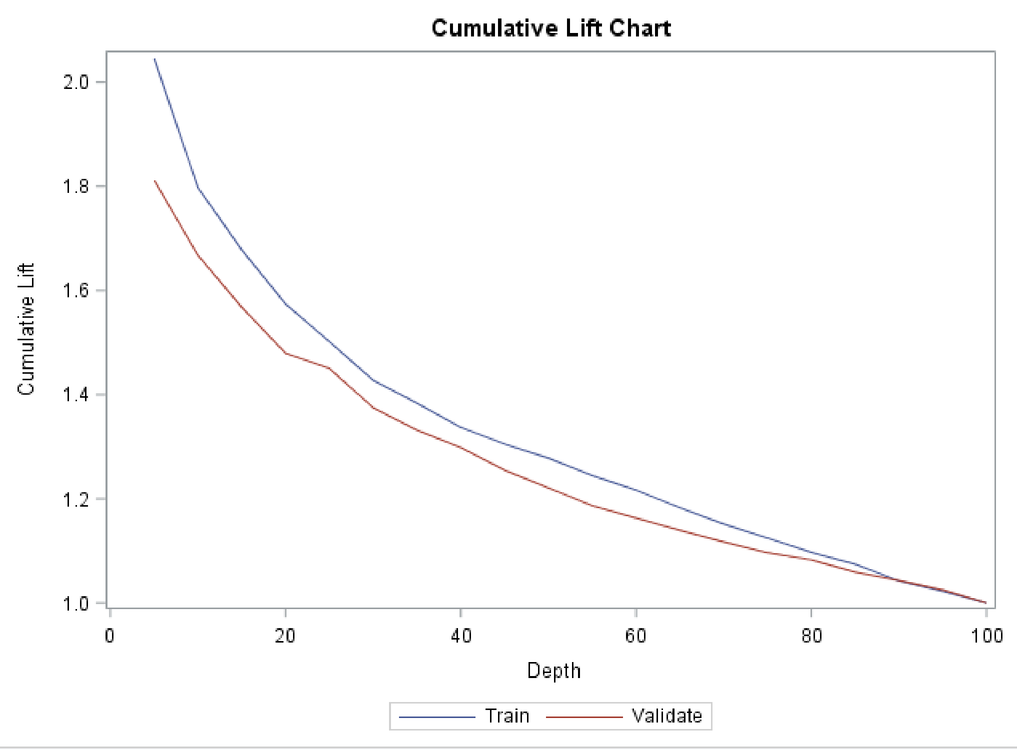
Misclassification and ROC Area Under the curve

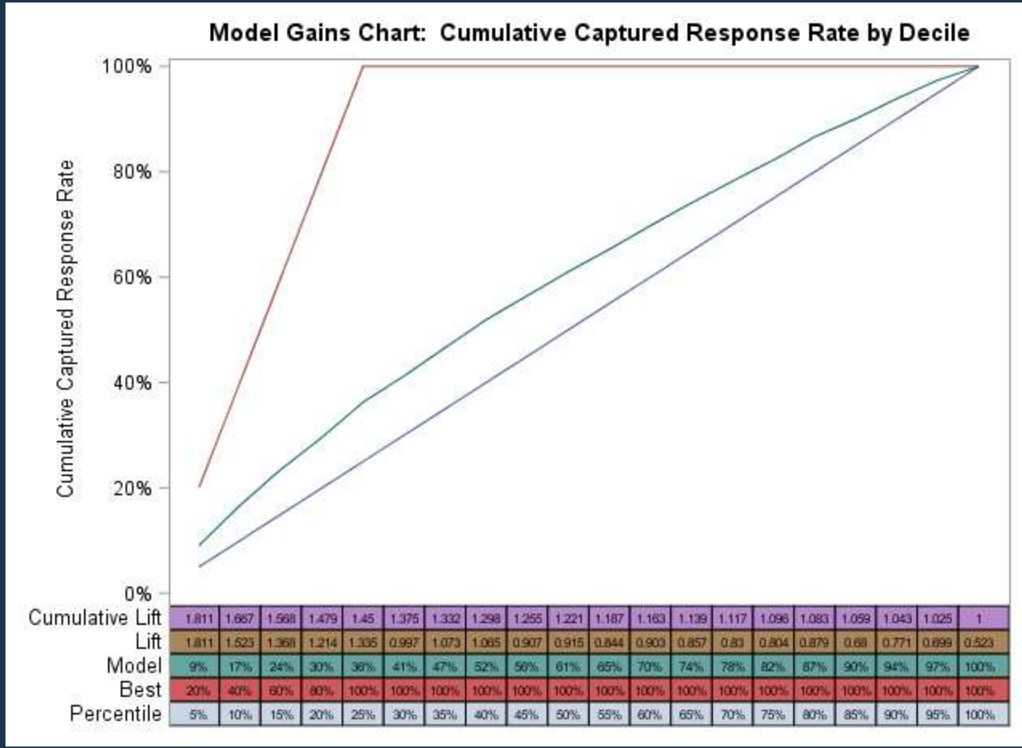
	# Terms	Misclassification	ROC
Model			
Full	38	0.2490	0.6229
Stepwise	13	0.2491	0.6204
Backward	13	0.2497	0.6205
Forward	13	0.2491	0.6204
*2-way Interactions	30	$\boxed{0.2465}$	$\boxed{0.6321}$
*2 degree Polynomial Terms	18	0.2491	$\boxed{0.6321}$
*2-Way + 2 degree Poly	31	0.2479	0.6314
		Smallest	Largest

* Using Forward Selection Method

SAS® Enterprise Guide®

2 Ways to Create Lift and Gains Charts





1. SAS Rapid Predictive Modeler (RPM) Task (need to license SAS Enterprise Miner to have access to this task)

SAS® Enterprise Guide®

2 Ways to Create Lift and Gains Charts

2. Through SAS Code

Sample *41683:* Gains and Lift plots for binary-response models

Details | Results | Downloads | About | Rate It

Gains and Lift plots for binary-response models

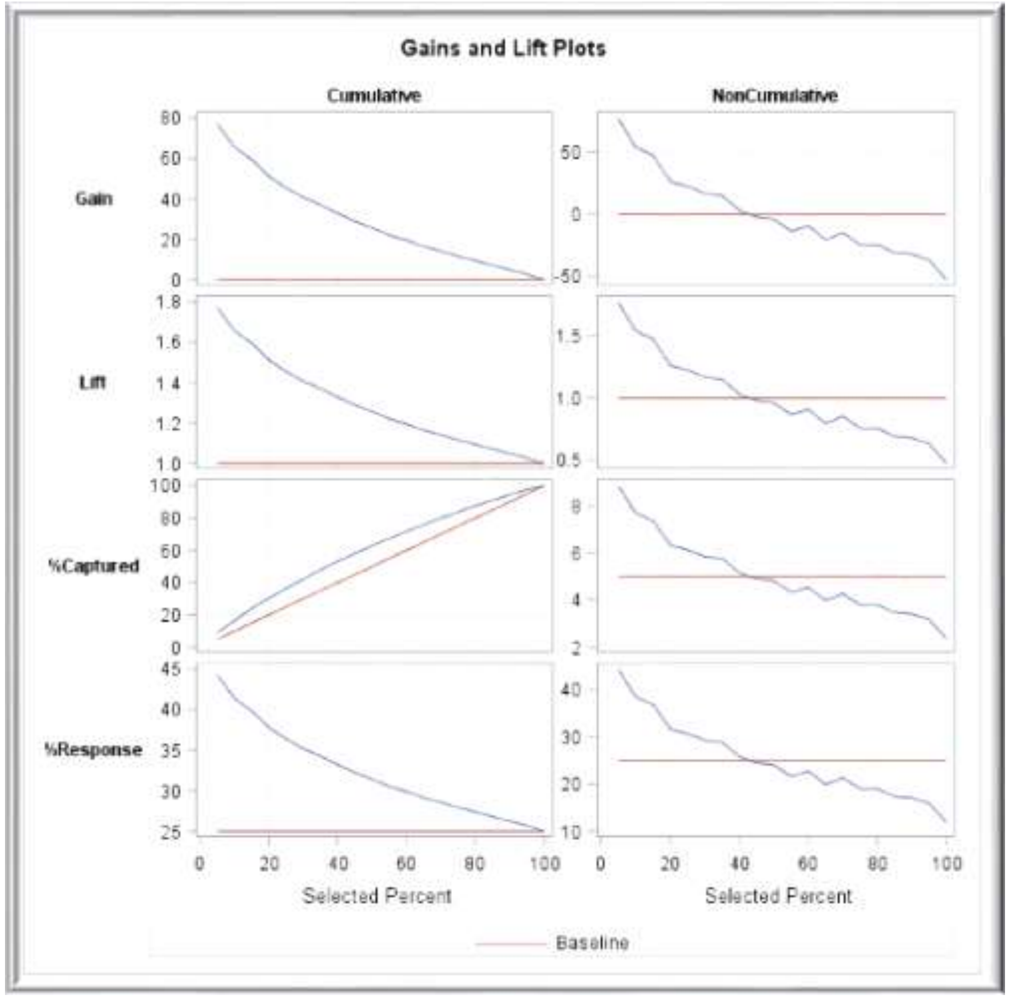
Contents: Purpose / History / Requirements / Usage / Details / Limitations / Missing Values

PURPOSE:

The %GainLift macro produces cumulative and noncumulative plots of the following statistics for a binary response model such as a logistic or probit model:

- Gains
- Lift
- Percent Captured
- Percent Response

All possible plots can be displayed in a single panel, or plots can be individually presented.

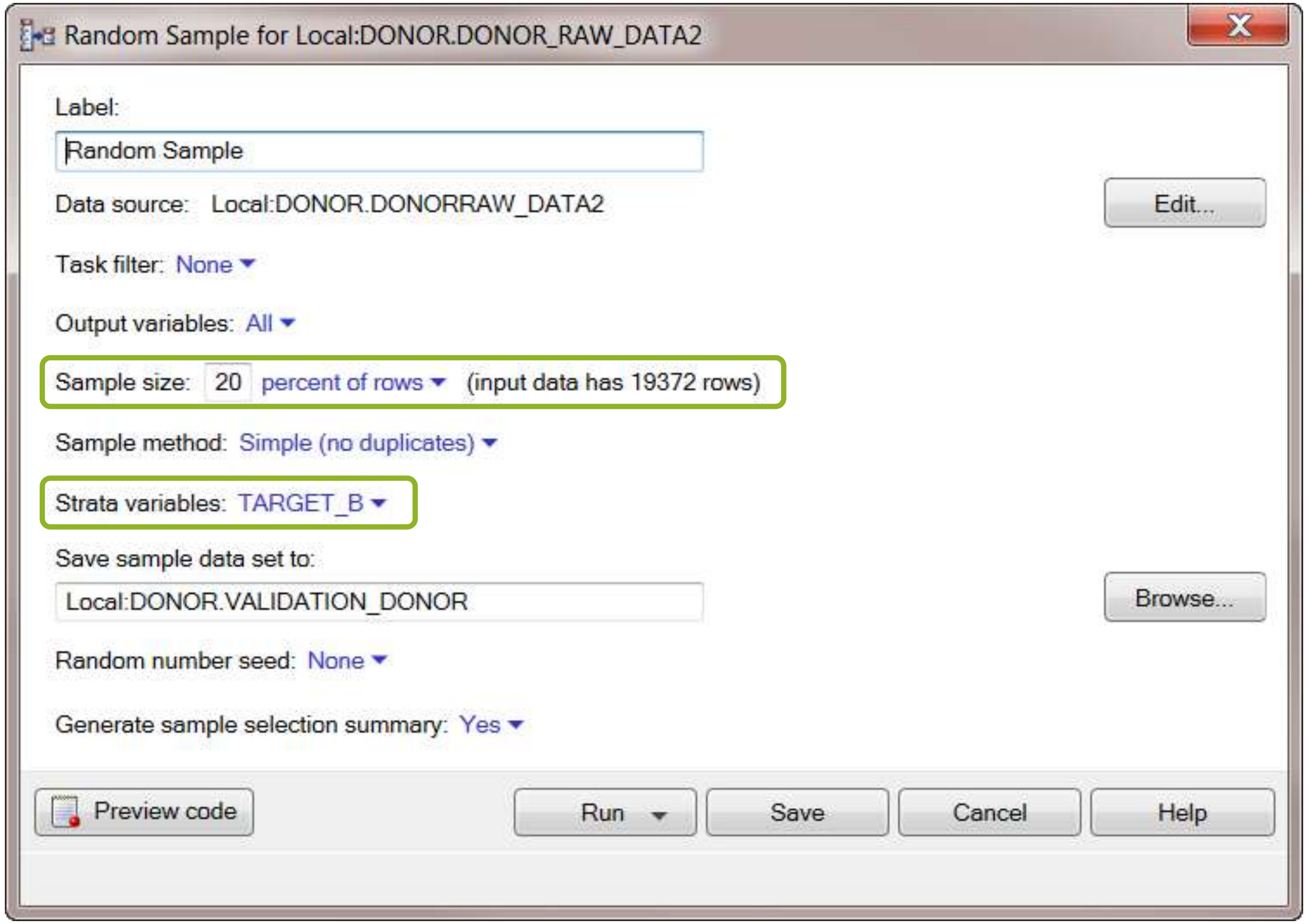


SAS® Enterprise Guide®

Honest Assessment

Create Validation Dataset

- Tasks → Data → Random Sample

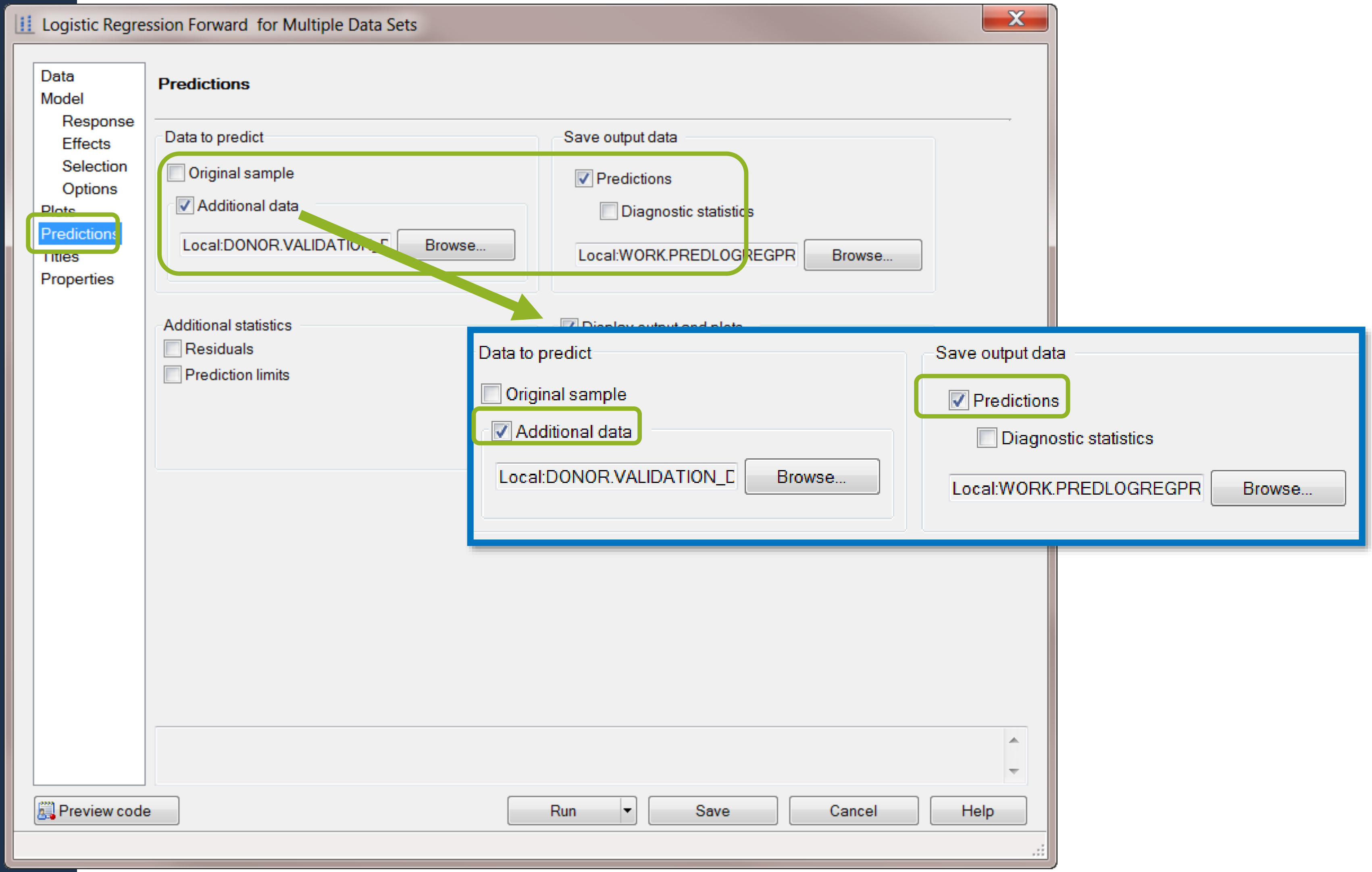


SAS® Enterprise Guide®

Misclassification

Honest Assessment

Method 1



SAS® Enterprise Guide®

Misclassification

```
proc logistic inmodel=donor.log_forw;
      score data=donor.validation_donor out=score1
      fitstat;
```

Honest Assessment

Method 2

Data Set	Total Frequency	Log Likelihood	Error Rate	AI
DONOR.VALIDATION_DONOR	3875	-2112.2	0.2483	4266.33

SAS® Enterprise Guide®

Validation Data

Honest Assessment

	# Terms	R-squared	Max Rescaled R2	AIC	SBC	Misclassification
Model						
Full	38	0.0344	0.0510	4322	4636	0.2488
Stepwise	13	0.0341	0.0504	4266	4397	0.2483
Backward	13	0.0338	0.0500	4265	$\boxed{4391}$	0.2501
Forward	13	0.0340	0.0504	4266	4398	0.2483
*2-way Interactions	30	0.0424	0.0629	4265	4496	$\boxed{0.2467}$
*2 degree Polynomial Terms	18	0.0380	0.0563	$\boxed{4252}$	4390	0.2475
*2-Way + 2 degree Poly	31	$\boxed{0.2634}$	$\boxed{0.3512}$	4263	4501	0.2477
		Largest	Largest	Smallest	Smallest	Smallest

* Using Forward Selection Method

SAS® Enterprise Guide®

Scoring in SAS analytical Procedures

Several Options depending on PROC. You can use a combination of

- OUTMODEL= and INMODEL= and SCORE
- OUTEST= and PROC SCORE
- STORE and PROC PML
- CODE

Introduction to Special SAS Data Sets

PROC SCORE

Techniques for scoring a regression model in SAS

SAS® Enterprise Guide®

Scoring in SAS *analytical Procedures*

CODE option

- Put this line of code after the model statement (;)
  - code file='c:\temp\logistic.sas';

```
*****************************************;
** SAS Scoring Code for PROC Logistic;
*****************************************;

length I_TARGET_B $ 12;
label I_TARGET_B = 'Into: TARGET_B' ;
label U_TARGET_B = 'Unnormalized Into: TARGET_B' ;

label P_TARGET_B1 = 'Predicted: TARGET_B=1' ;
label P_TARGET_B0 = 'Predicted: TARGET_B=0' ;

drop _LMR_BAD;
_LMR_BAD=0;

*** Check interval variables for missing values;
if nmiss(IN_HOUSE,MEDIAN_HOME_VALUE,PEP_STAR,FREQUENCY_STATUS_97NK,
        RECENT_AVG_GIFT_AMT,RECENT_CARD_RESPONSE_PROP,LIFETIME_CARD_PROM,
        LIFETIME_MIN_GIFT_AMT,NUMBER_PROM_12,MONTHS_SINCE_LAST_GIFT,
        MONTHS_SINCE_FIRST_GIFT,IM_WEALTH_RATING) then do;
   _LMR_BAD=1;
   goto _SKIP_000;
end;
```

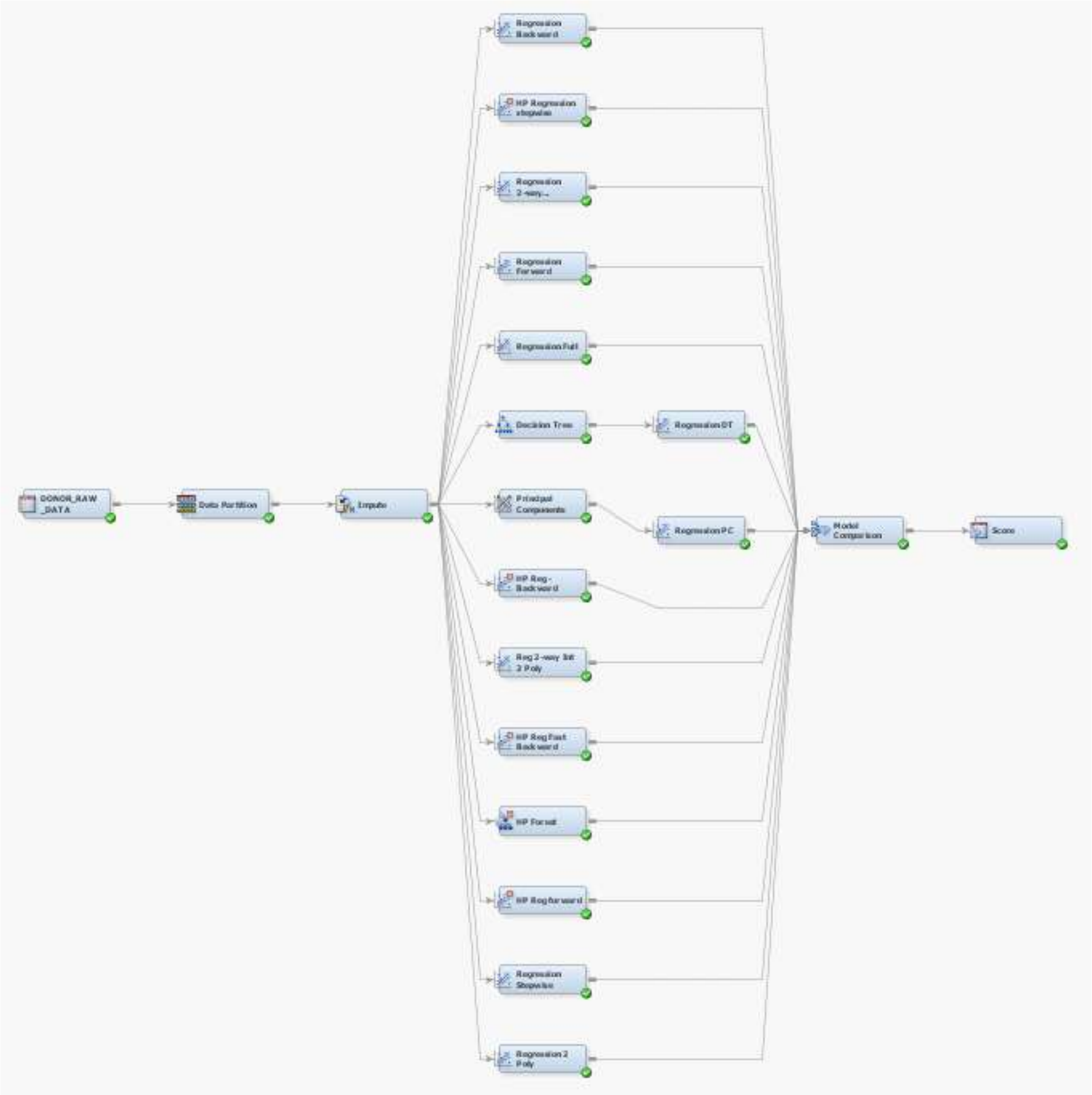
Top of Scoring Code

Lots more code in between

```
select( _IY );
   when (1) do;
      I_TARGET_B = '1' ;
      U_TARGET_B = 1;
   end;
   when (2) do;
      I_TARGET_B = '0' ;
      U_TARGET_B = 0;
   end;
   otherwise do;
      I_TARGET_B = '';
      U_TARGET_B = .;
   end;
end;
_SKIP_000:
if _LMR_BAD = 1 then do;
I_TARGET_B = '';
U_TARGET_B = .;
P_TARGET_B1 = .;
P_TARGET_B0 = .;
end;
drop _TEMP;
```

Bottom of Scoring Code

SAS® Enterprise Miner™

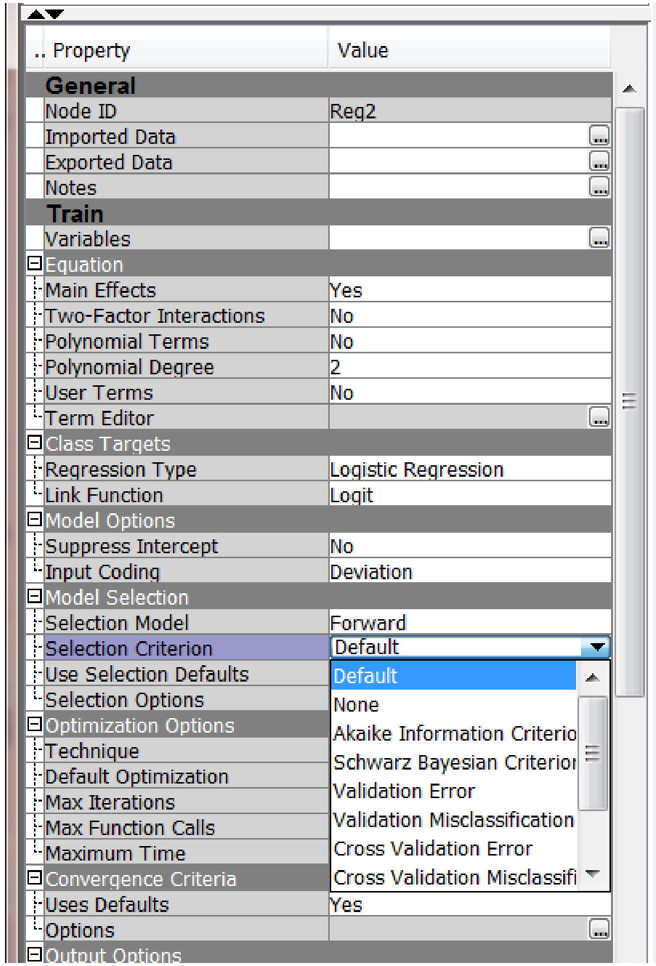


SAS Enterprise Miner’s strength is the ability to create many models with different settings or different modeling techniques and compare all to determine the winning model.

SAS® Enterprise Miner™

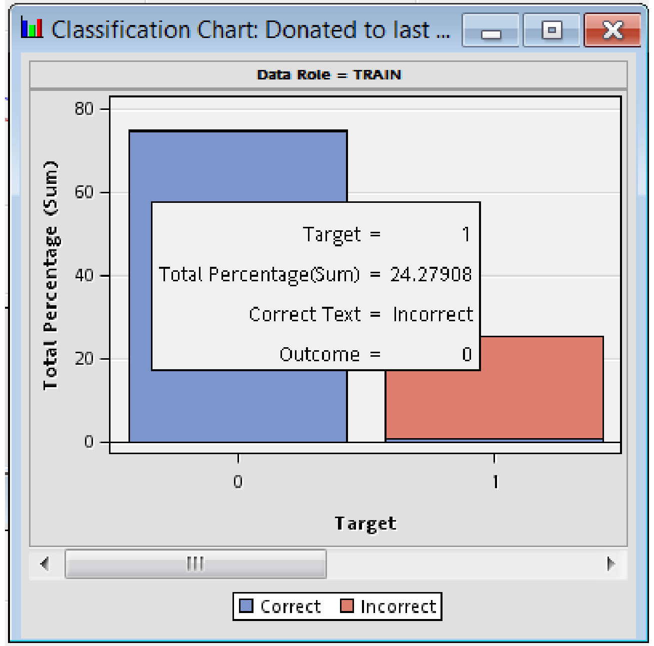
Regression – many criteria available for Model selection

- AIC
- SB
- Validation Error
- Validation Misclassification
- Cross Validation Error
- Cross Validation Misclassification
- Profit/Loss
- Validation Profit /Loss
- Cross Validation Profit/Loss



SAS® Enterprise Miner™

Regression Output



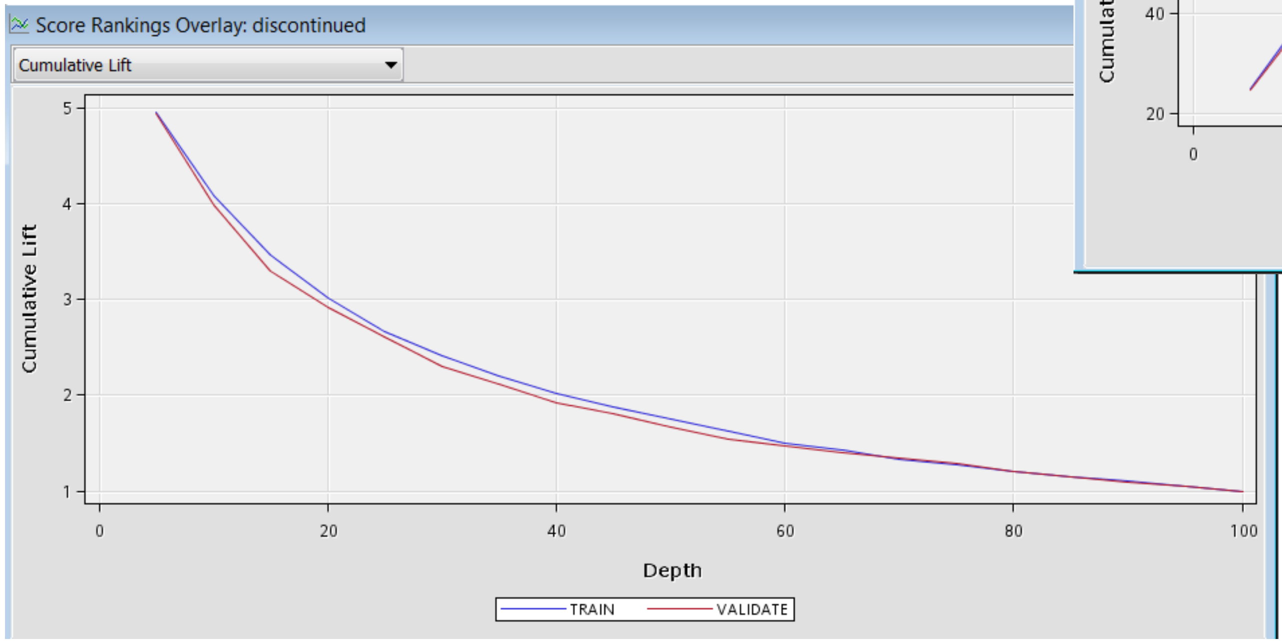
Fit Statistics

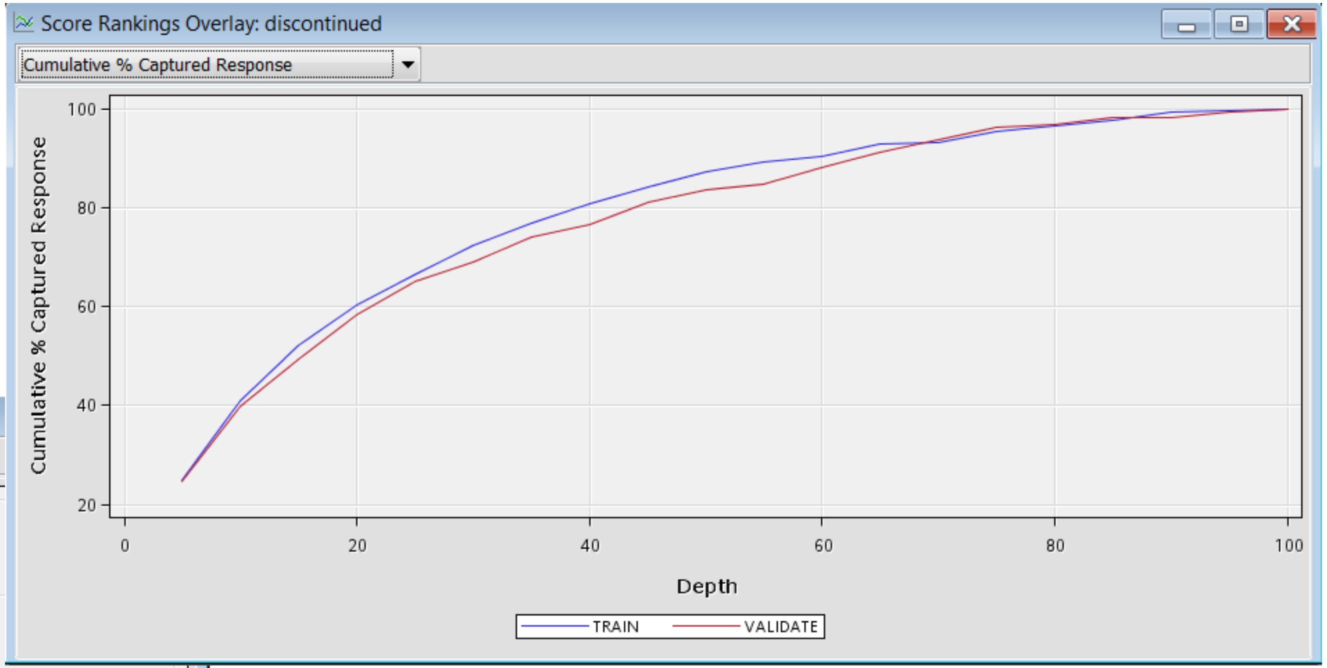
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
TARGET_B	Donated to last ca...	AIC	Akaike's Informati...	14767.72	.
TARGET_B	Donated to last ca...	ASE	Average Squared ...	0.179434	0.182505
TARGET_B	Donated to last ca...	AVERR	Average Error Fun...	0.541549	0.550794
TARGET_B	Donated to last ca...	DFE	Degrees of Freed...	13518	.
TARGET_B	Donated to last ca...	DFM	Model Degrees of ...	41	.
TARGET_B	Donated to last ca...	DFT	Total Degrees of ...	13559	.
TARGET_B	Donated to last ca...	DIV	Divisor for ASE	27118	11626
TARGET_B	Donated to last ca...	ERR	Error Function	14685.72	6403.535
TARGET_B	Donated to last ca...	FPE	Final Prediction Er...	0.180522	.
TARGET_B	Donated to last ca...	MAX	Maximum Absolut...	0.904725	0.999651
TARGET_B	Donated to last ca...	MSE	Mean Square Error	0.179978	0.182505
TARGET_B	Donated to last ca...	NOBS	Sum of Frequencies	13559	5813
TARGET_B	Donated to last ca...	NW	Number of Estima...	41	.
TARGET_B	Donated to last ca...	RASE	Root Average Su...	0.423596	0.427206
TARGET_B	Donated to last ca...	RFPE	Root Final Predicti...	0.424879	.
TARGET_B	Donated to last ca...	RMSE	Root Mean Squar...	0.424238	0.427206
TARGET_B	Donated to last ca...	SBC	Schwarz's Bayesi...	15075.82	.
TARGET_B	Donated to last ca...	SSE	Sum of Squared E...	4865.888	2121.8
TARGET_B	Donated to last ca...	SUMW	Sum of Case Wei...	27118	11626
TARGET_B	Donated to last ca...	MISC	Misclassification ...	0.247585	0.251161

SAS® Enterprise Miner™

Regression Output

Cumulative Lift



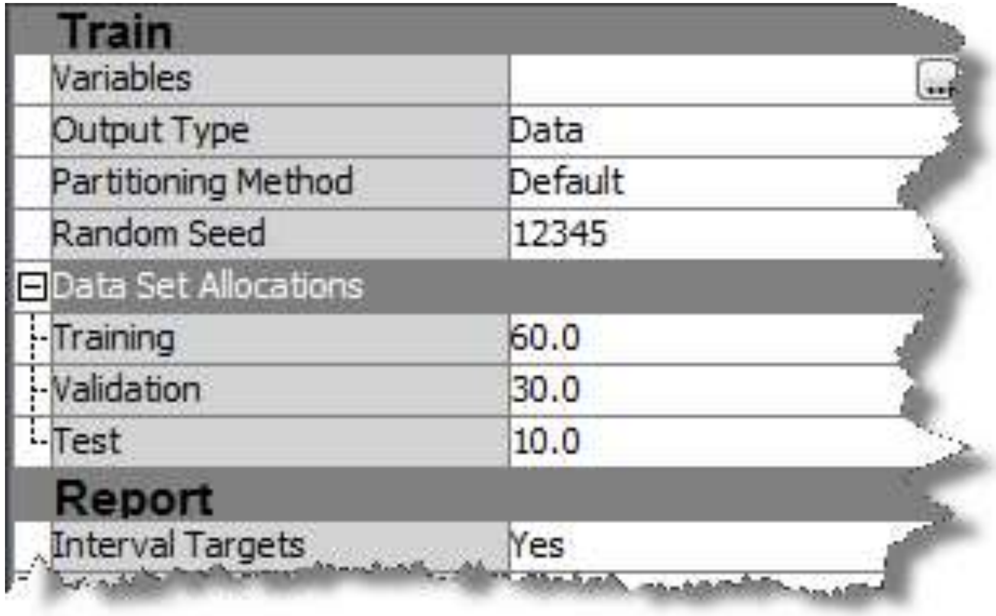


Cumulative Gains

Honest Assessment

Data Partition Node

Data Mining Best Practice of Partitioning Data into training, validation and test data sets incorporated

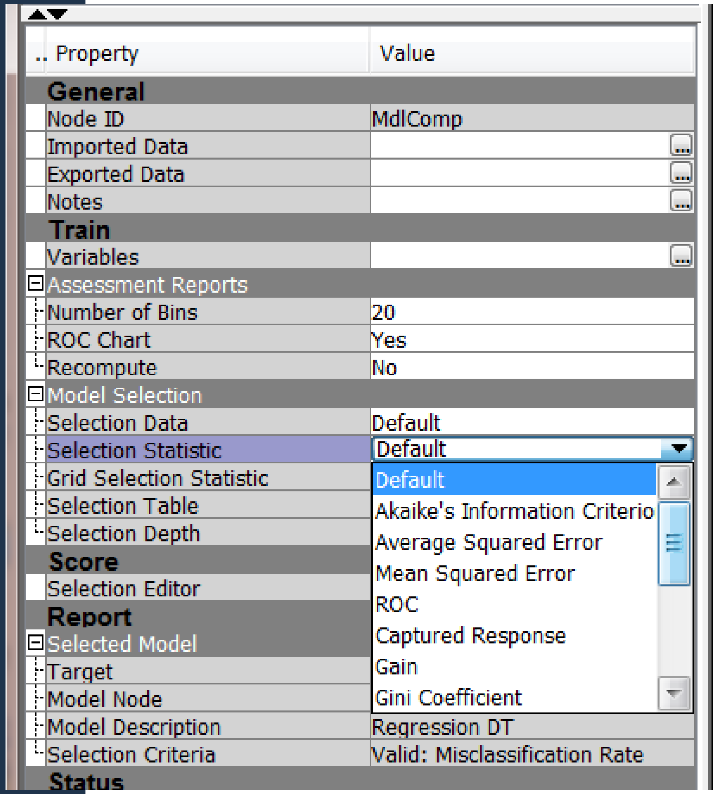


Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	60.0
Validation	30.0
Test	10.0
Report	
Interval Targets	Yes

Including the ability to partition using Simple Random, Clustering or Stratified methods

SAS® Enterprise Miner™

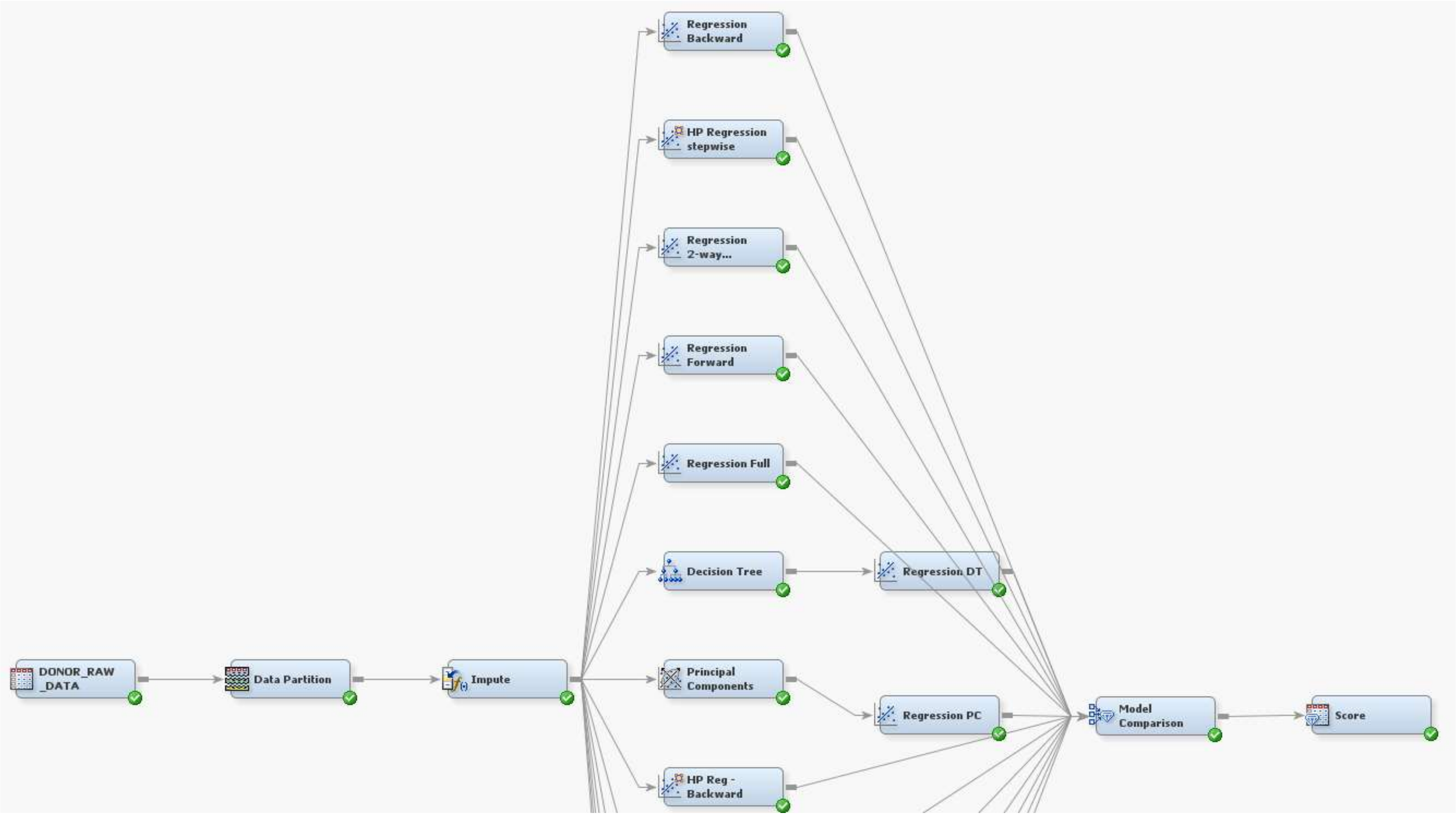
Model Comparison Node

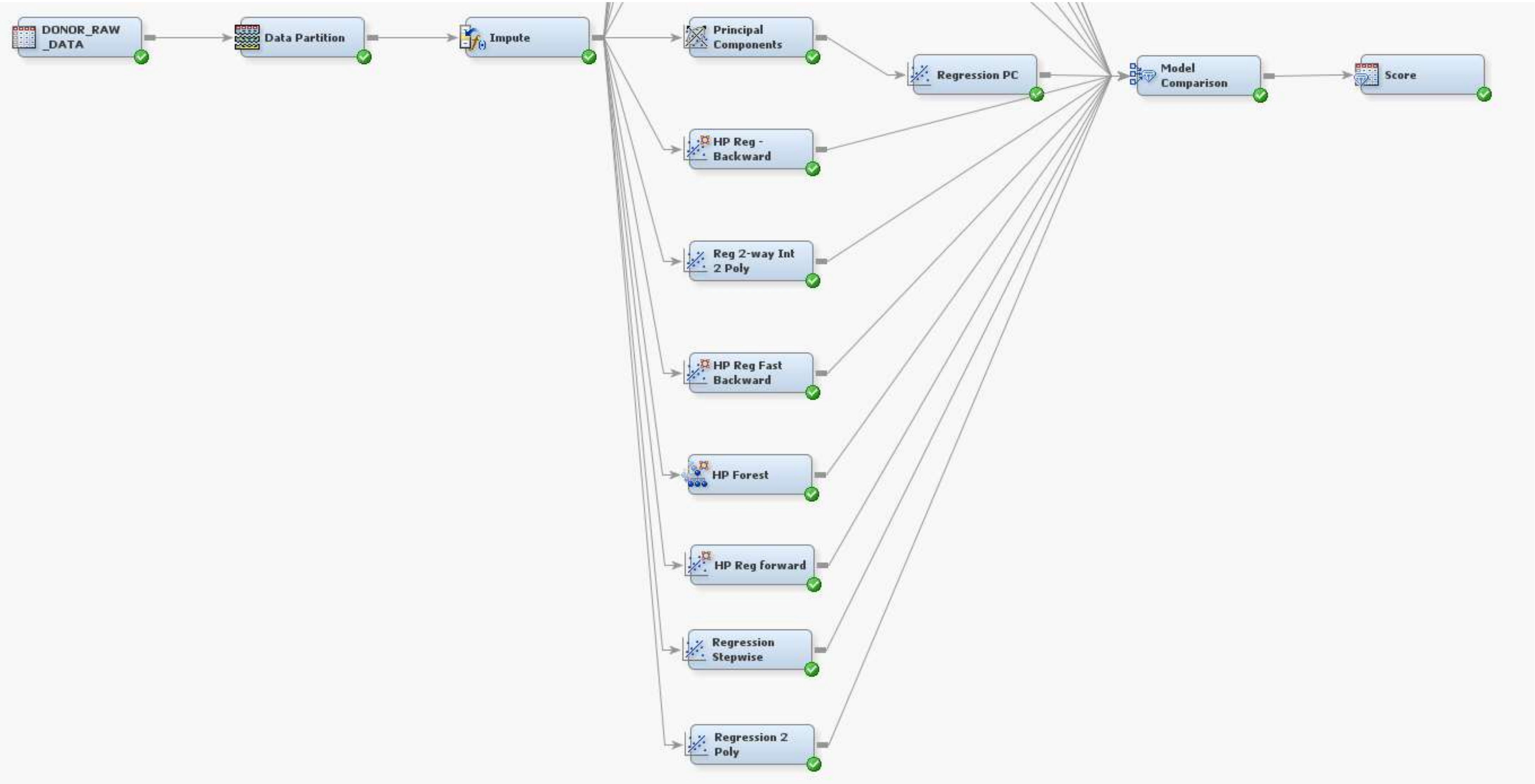


The Model Comparison node provides a common framework for comparing models and predictions from any of the modeling tools (such as Regression, Decision Tree, and Neural Network tools). The comparison is based on standard model fits statistics as well as potential expected and actual profits or losses that would result from implementing the model. The node produces the following charts that help to describe the usefulness of the model: lift, profit, return on investment, receiver operating curves, diagnostic charts, and threshold-based charts.

AIC	Captured Response
ASE	KS Statistic
MSE	Misclassification
ROC	Average Profit/Loss
Gain	Cumulative Lift
Lift	Cumulative Captured Response
Gini	Cumulative Percent Response

Available for training, validation and test datasets


Regression Backward
HP Regression stepwise
Regression 2-way...
Regression Forward
Regression Full
Decision Tree
Regression DT
DONOR_RAW _DATA
Data Partition
Impute
Principal Components
Regression PC
Model Comparison
Score
HP Reg - Backward


DONOR_RAW
_DATA
Data Partition
Impute
Principal
Components
Regression PC
Model
Comparison
Score
HP Reg -
Backward
Reg 2-way Int
2 Poly
HP Reg Fast
Backward
HP Forest
HP Reg forward
Regression
Stepwise
Regression 2
Poly

SAS® Enterprise Miner™
Model Comparison Node

Best Model

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate ▲	Train: Misclassification Rate	Valid: Lift	Train: Schwarz's Bayesian Criterion
Y	Reg4	Reg4	Regression DT	TARGET...	Donated ...	0.249441	0.24965	1.539784	15059.33
	HPDMFo...	HPDMFo...	HP Forest	TARGET...	Donated ...	0.249957	0.249797	1.429799	.
	HPReg4	HPReg4	HP Regression stepwise	TARGET...	Donated ...	0.250473	0.249428	1.546658	.
	Reg5	Reg5	Regression PC	TARGET...	Donated ...	0.250645	0.249133	1.443547	14993.11
	HPReg	HPReg	HP Reg - Backward	TARGET...	Donated ...	0.250817	0.249281	1.457295	.
	HPReg3	HPReg3	HP Reg forward	TARGET...	Donated ...	0.250989	0.247585	1.374807	.
	Reg2	Reg2	Regression Forward	TARGET...	Donated ...	0.251161	0.247585	1.361059	15075.82
	Reg3	Reg3	Regression Stepwise	TARGET...	Donated ...	0.251161	0.247585	1.361059	15075.82
	Reg	Reg	Regression Backward	TARGET...	Donated ...	0.251849	0.247732	1.361059	15075.56
	HPReg2	HPReg2	HP Reg Fast Backward	TARGET...	Donated ...	0.252193	0.248838	1.539784	.
	Reg8	Reg8	Regression 2 Poly	TARGET...	Donated ...	0.253226	0.246478	1.484792	15017.38
	Reg6	Reg6	Regression Full	TARGET...	Donated ...	0.253398	0.246773	1.622272	15639.84
	Reg9	Reg9	Reg 2-way Int 2 Poly	TARGET...	Donated ...	0.258214	0.241463	1.429799	16427.17
	Reg7	Reg7	Regression 2-way Interactions	TARGET...	Donated ...	0.295544	0.21211	1.127342	33523.52

SAS Enterprise Miner assumes decision processing and selects the model with the lowest misclassification rate when there is a binary target.

Which?

Model Assessment

Criterion

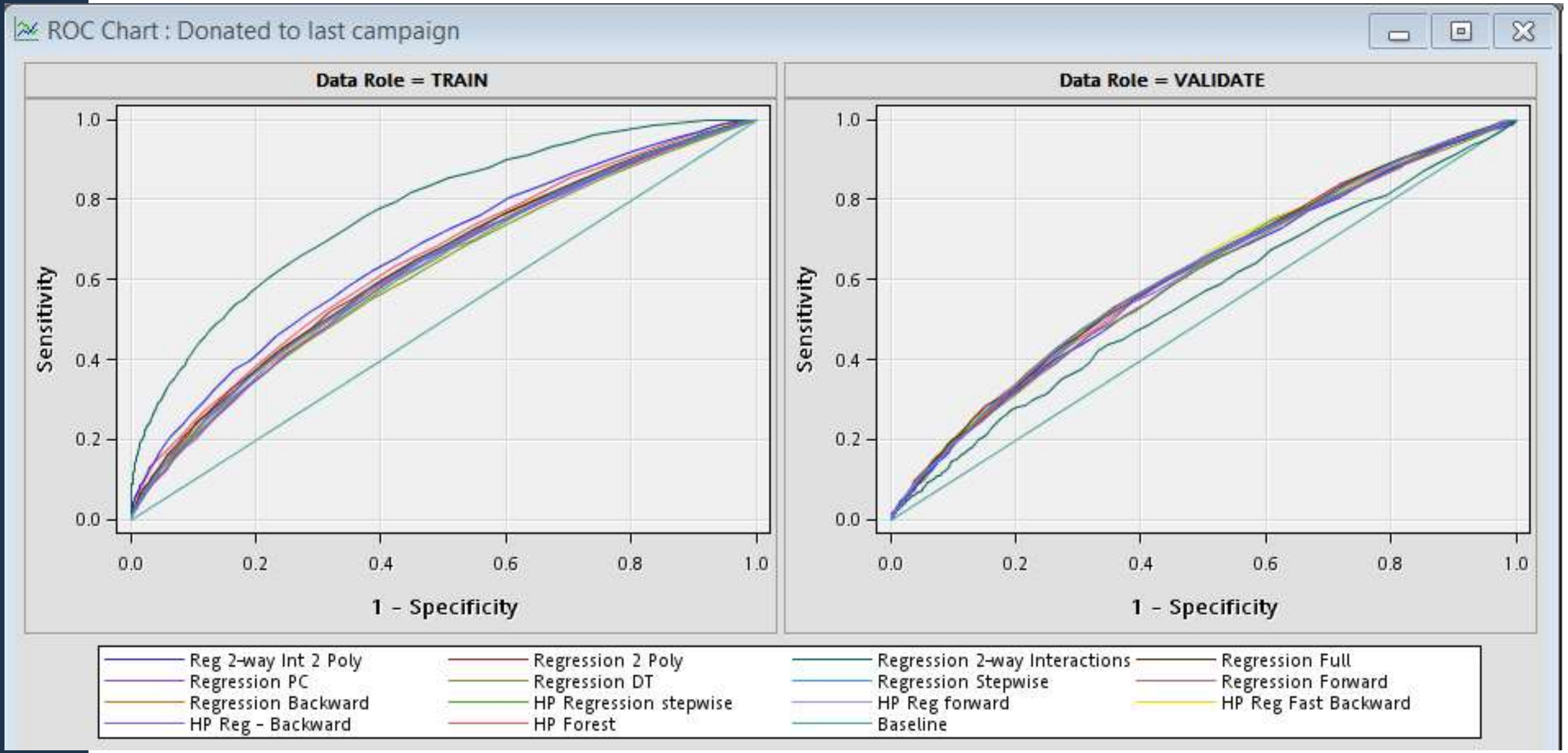
- For decision prediction
  - Accuracy & Misclassification
  - Profit or loss
  - Kolmogorov-Smirnov (KS) Statistic
- For ranking predictions
  - ROC index
  - GINI coefficient
- For estimate predictions
  - Akaike information criterion (AIC)
  - Schwarz’s Bayesian Criterion (SBC)
  - Average squared error

[Defining Measures of Success for Predictive Models]

SAS Enterprise Miner Help under Model Comparison for additional information

SAS® Enterprise Miner™

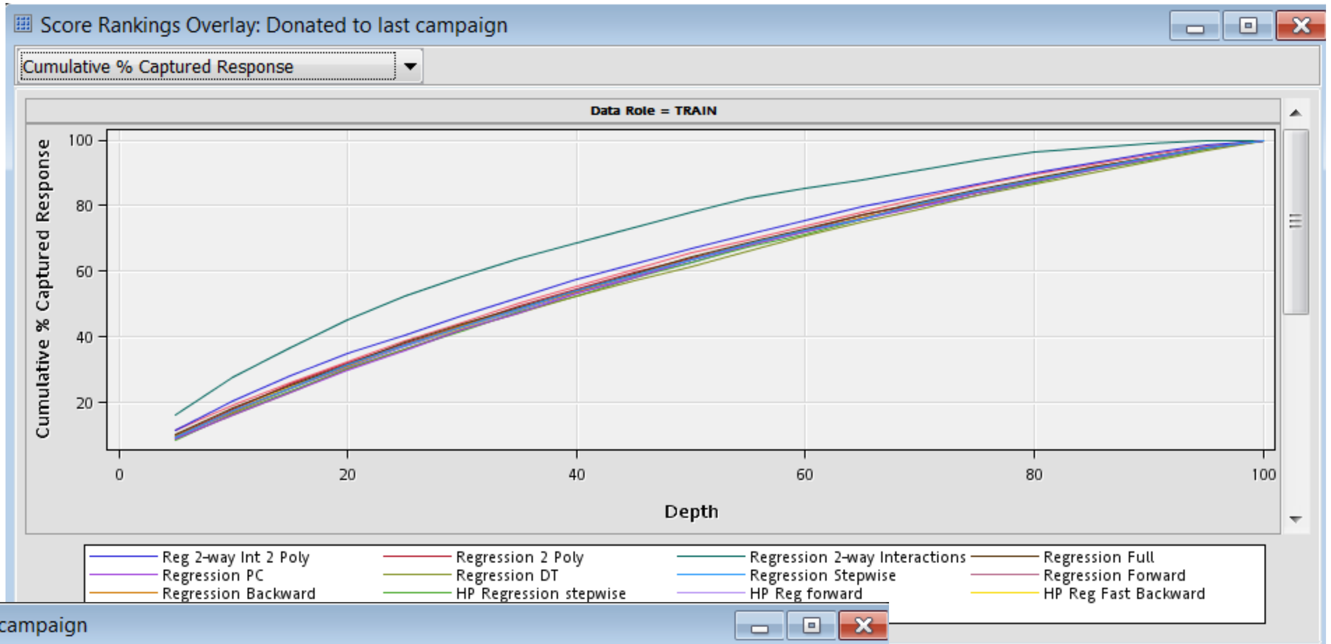
ROC Chart

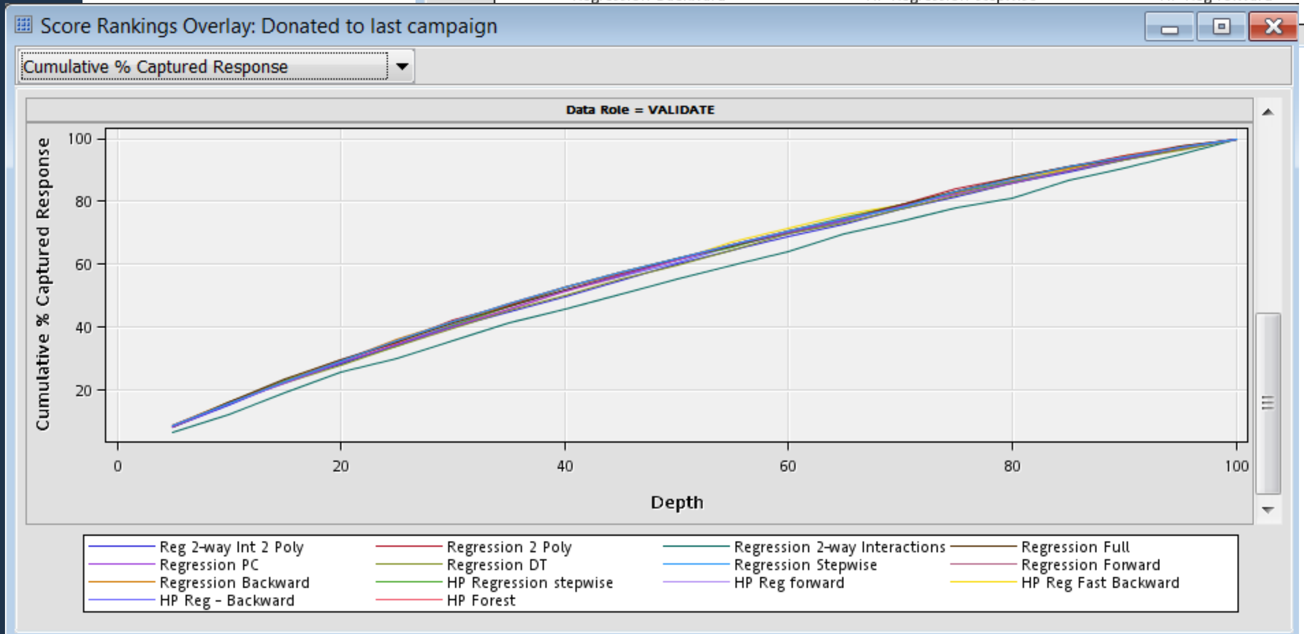


SAS® Enterprise Miner™

Cumulative Gains Chart

Train



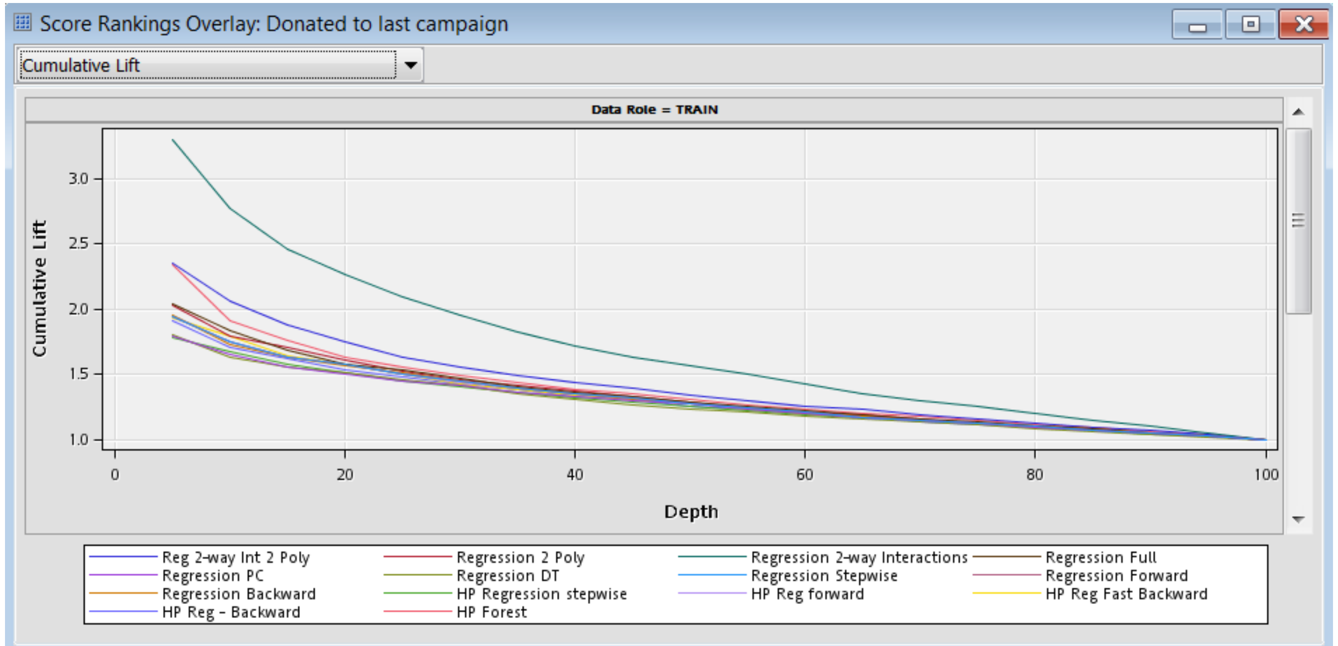


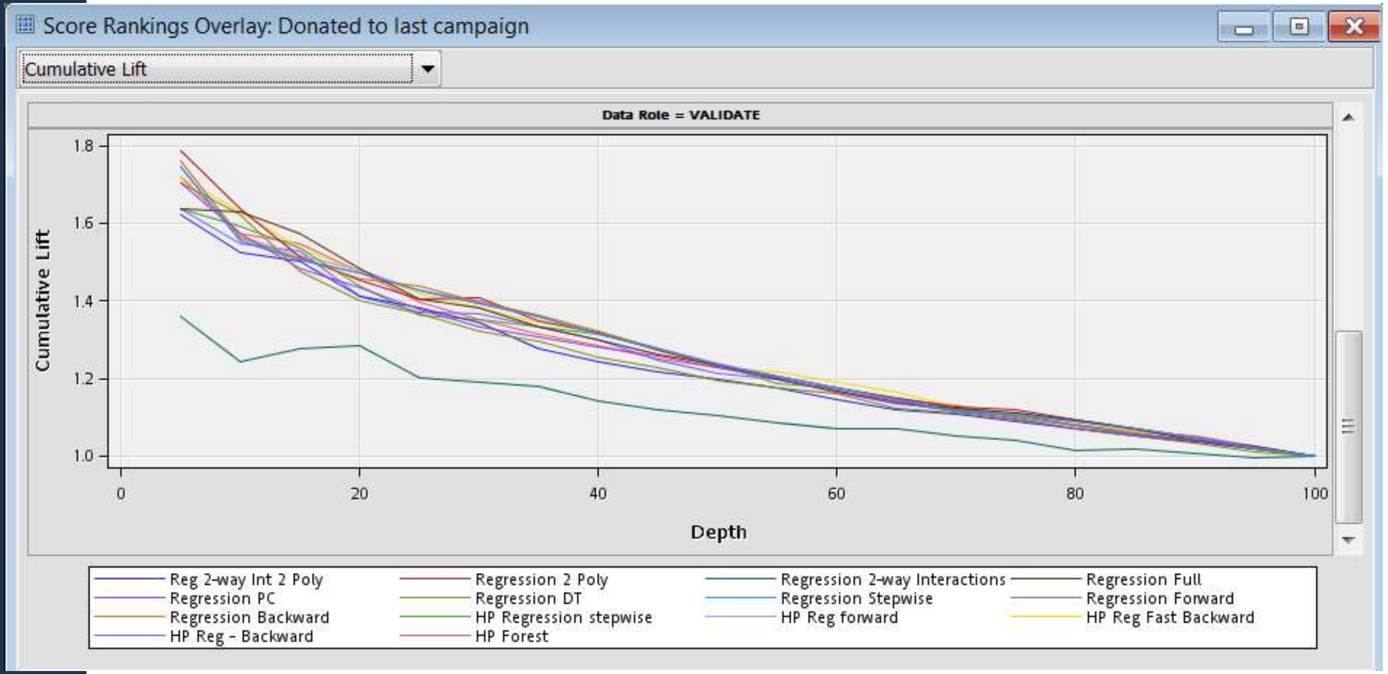
Validation

SAS® Enterprise Miner™

Cumulative Lift Chart

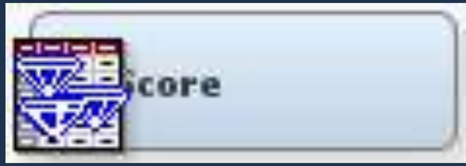
Train





Validation

Score Node



Property	Value
General	
Node ID	Score
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Type of Scored Data	View
Use Fixed Output Names	Yes
Hide Variables	No
Hide Selection	...
Score Data	
Validation	No
Test	No
Score Code Generation	
Optimized Code	Yes
C Score	No
Java Score	No
Java Package Name	Default
User Package Name	
Report	
Graphical Reports	Yes
Status	

The Score node enables you to manage, edit, export, and execute scoring code that is generated from a trained model. Scoring is the generation of predicted values for a data set that may not contain a target variable. The Score node generates and manages scoring formulas in the form of a single SAS DATA step, which can be used in most SAS environments even without the presence of Enterprise Miner.

SAS® Enterprise Miner™

Optimized Score Code from Score Node

Optimized to only include variables in the final model

```
***************************************;
*** begin scoring code for regression;
***************************************;

length _WARN_ $4;
label _WARN_ = 'Warnings' ;

length I_TARGET_B $ 12;
label I_TARGET_B = 'Into: TARGET_B' ;
*** Target Values;
array REG4DRF [2] $12 _temporary_ ('1' '0' );
label U_TARGET_B = 'Unnormalized Into: TARGET_B' ;
*** Unnormalized target values;
ARRAY REG4DRU[2]  _TEMPORARY_ (1 0);

drop _DM_BAD;
_DM_BAD=0;

*** Check FILE_CARD_GIFT for missing values ;
if missing( FILE_CARD_GIFT ) then do;
   substr(_warn_,1,1) = 'M';
   _DM_BAD = 1;
end;
```

Top of Scoring Code

Lots more code in between

```
*------------------------------------------------------------*;
* TOOL: Score Node;
* TYPE: ASSESS;
* NODE: Score;
*------------------------------------------------------------*;
*------------------------------------------------------------*;
* Score: Creating Fixed Names;
*------------------------------------------------------------*;
LABEL EM_SEGMENT = 'Segment';
EM_SEGMENT = b_TARGET_B;
LABEL EM_EVENTPROBABILITY = 'Probability for level 1 of TARGET_B';
EM_EVENTPROBABILITY = P_TARGET_B1;
LABEL EM_PROBABILITY = 'Probability of Classification';
EM_PROBABILITY =
max(
P_TARGET_B1
,
P_TARGET_B0
);
LENGTH EM_CLASSIFICATION $%dmnorlen;
LABEL EM_CLASSIFICATION = "Prediction for TARGET_B";
EM_CLASSIFICATION = I_TARGET_B;
```

Bottom of Scoring Code

Resources for Model Selection

Resources
Model Selection

Defining Measures of Success for Predictive Models

- http://www.predictiveanalyticsworld.com/patimes/defining-measures-of-success-for-predictive-models-0608152/5519/

The Evolution of Analytics: Opportunities and Challenges for Machine Learning in Business

- http://resources.cio.com/ccd/assets/110448/detail

An Empirical Comparison of Supervised Learning Algorithms

- http://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf

Least Angle Regression

- http://statweb.stanford.edu/~imj/WEBLIST/2004/LarsAnnStat04.pdf

The Analysis and Selection of Variables in Linear Regression

- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.472.4742&rep=rep1&type=pdf

Resources

Additional Reading for SAS

Introducing the GLMSELECT PROCEDURE for Model Selection

- http://www2.sas.com/proceedings/sugi31/207-31.pdf

Model Selection in Linear Mixed Effects Models Using SAS PROC MIXED

- http://www2.sas.com/proceedings/sugi22/STATS/PAPER284.PDF

SAS Code to Select the Best Multiple Linear Regression Model for Multivariate Data Using Information Criteria

- http://analytics.ncsu.edu/sesug/2005/SA01_05.PDF

Recreating the SELECTION=SCORE Model Specification with the BEST=n Effect Selection Option for PROC SURVEYLOGISTIC

- http://www.lexjansen.com/wuss/2007/AnalyticsStatistics/ANL_Adams_RecreatingSelection.pdf

Gentle Introduction to Information Theoretic Model Selection Criteria and Their Applications in Clinical Trial

- http://www.lexjansen.com/nesug/nesug97/posters/chen.pdf

The Steps to Follow in a Multiple Regression Analysis

- http://support.sas.com/resources/papers/proceedings12/333-2012.pdf

Resources
SAS Courses

- Statistics 1: Introduction to ANOVA, Regression and Logistic Regression
- Statistics 2: ANOVA and Regression
- Categorical Data Analysis Using Logistic Regression
- Predictive Modeling Using SAS High-Performance Analytics Procedures
- Predictive Modeling Using Logistic Regression
- Applied Analytics Using SAS Enterprise Miner
- Data Mining: Principles and Best Practices

For a complete list of courses, please see
https://support.sas.com/edu/courses.html?ctry=us

Resources

Videos

- The HPBIN Procedure
- Introducing the HPGENSELECT Procedure
- Introducing PROC QUANTSELECT
- Fitting a Multiple Linear Regression Model with Stepwise Selection
- What’s New in SAS Enterprise Miner
- Interval Target Scorecards – Interactive Binning Node
- The New HP GLM Node in SAS Enterprise Miner
- Tutorials for SAS programming, Enterprise Guide, Analytics

Questions?

Thank you for your time and attention!

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

sas.com