Predictive Modeling with SAS (for Health)

Lorne Rothman, PhD, P.Stat.
Principal Analytics Services
Lorne.Rothman@sas.com

Overview

- What is Predictive Modeling?
  - Purpose, challenges, and methods
  - Examples
  - The North Carolina Low Birth Weight Data
- Data & Variable Preparation
  - Oversampling, Missing Values, Data splitting & dimension reduction
- Binary Target Modeling
  - Decision Trees with HPSPLIT
  - Logistic Regression with LOGISTIC
  - Comparing ROC curves
- Continuous Target Modeling
  - Model selection with GLMSELECT
- Generalized Linear Predictive Modeling
  - Gamma regression model selection with HPGENSELECT

What is Predicted Modeling?

THE POWER TO KNOW.®

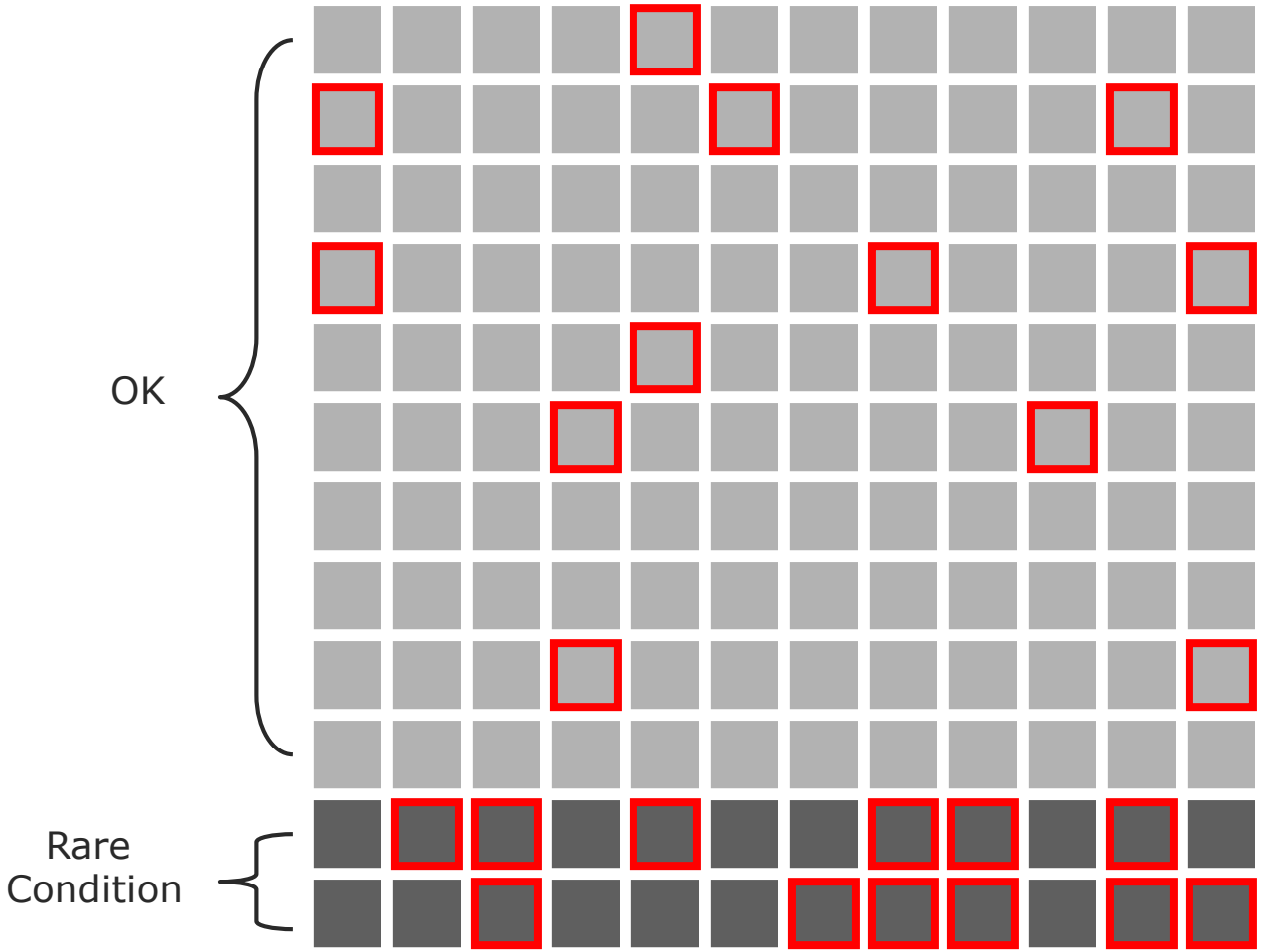
Purpose of Predictive Modeling

- ✓ To Predict the Future
- X To identify statistically significant attributes or risk factors
- X To publish findings in Science, Nature, or the New England Journal of Medicine
- ✓ To enhance & enable rapid decision making at the level of the individual patient, client, customer, etc.
- X To enable decision making and influence policy through publications and presentations

Data Deluge

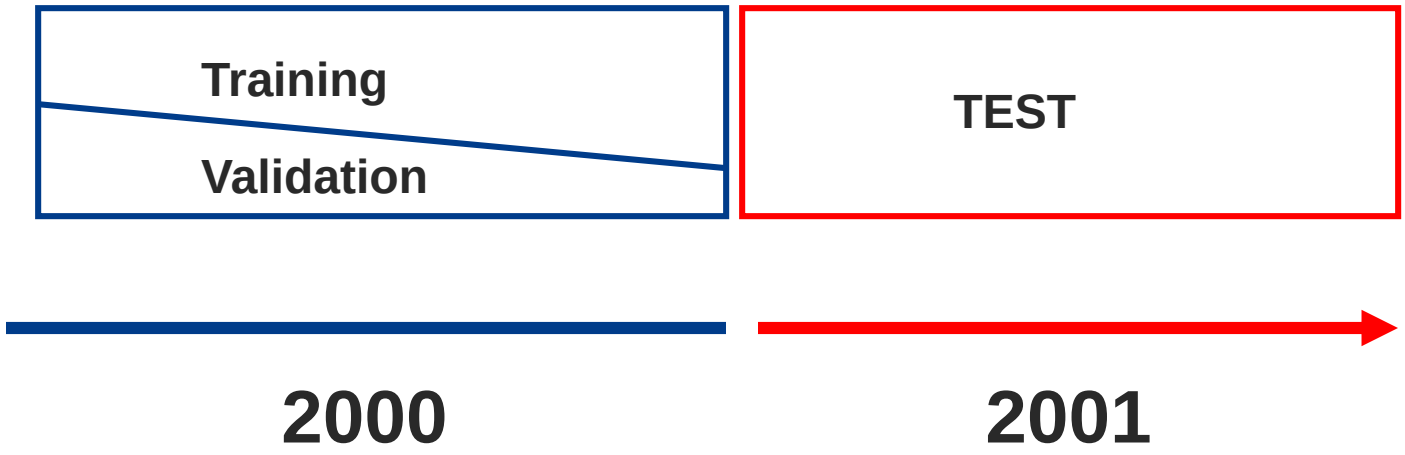
OHIP CIHI Population Census Canadian Cancer Registry Canadian Community Health Survey Canadian MIS Database Canadian Joint Replacement Registry Canadian Organ Replacement Registry Canadian Tobacco Use Monitoring Survey Discharge Abstract Database Disease Surveillance General Social Survey – Cycle 6 Health and Activity Limitation Survey Health Services Access Survey Hospital Mental Health Survey Hospital Morbidity Database Maternity Experiences Survey Natchem/Particulate Matter Database National Alcohol and Drug Survey Database National Pollutant Release Inventory National Population Health Survey (Household) National Rehabilitation Reporting System National Trauma Registry Survey on Aging and Independence Survey on Smoking in Canada TRAID Vital Statistics

Challenges: Rare Events



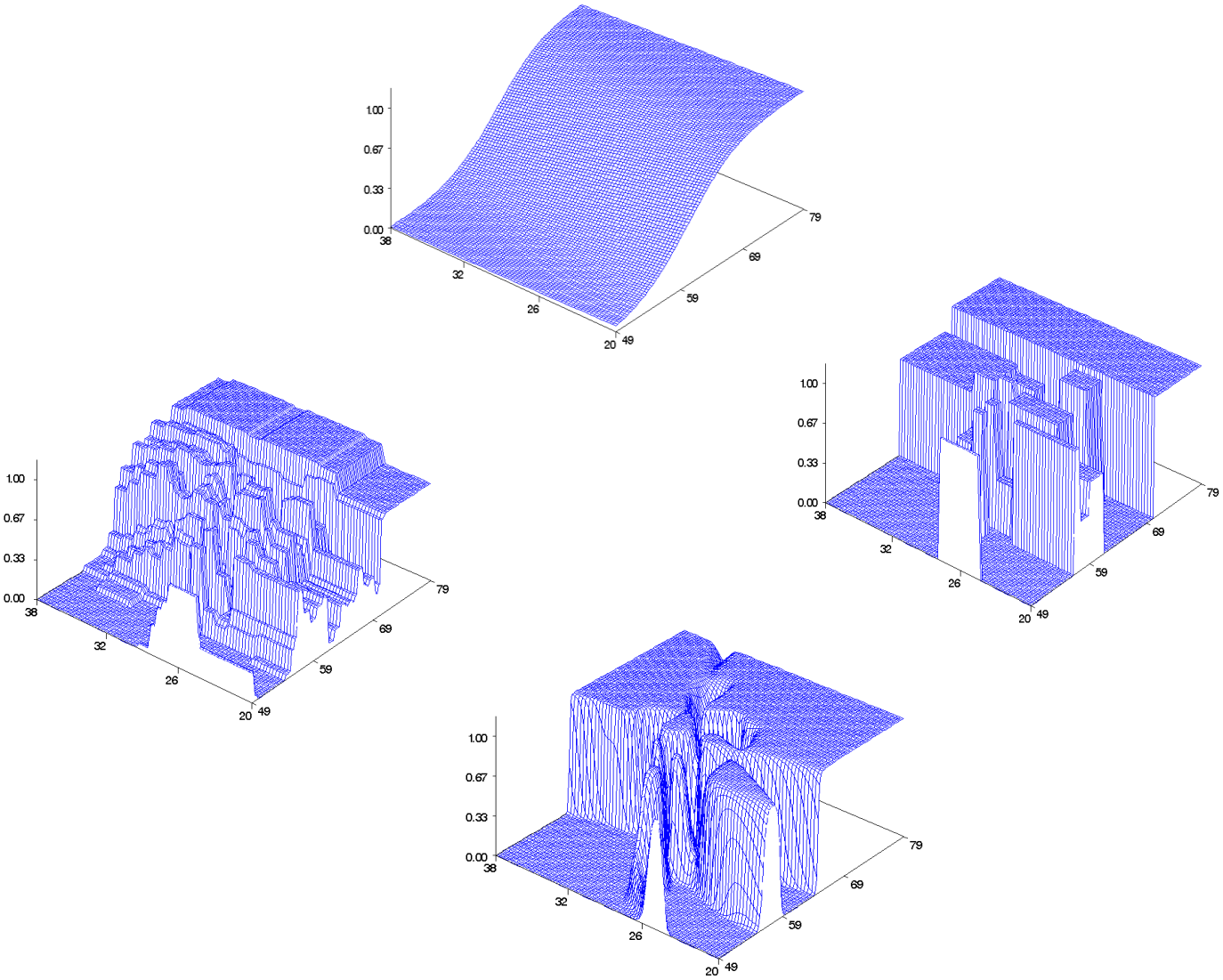
Methodology: Empirical Validation

Predicting the Future with Data Splitting



- Models are fit to Training Data, compared and selected on Validation and tested on a future Test set.

Methodology: Diversity of Algorithms



Jargon…

- Target = Dependent Variable.
- Inputs, Predictors = Independent Variables.
- Supervised Classification = Predicting class membership with algorithms that use a target.
- Scoring = The process of generating predictions on new data for decision making. This is not a re-running of models but an application of model results (e.g. equation and parameter estimates) to new data.
- Scoring Code = programming code that can be used to prepare and generate predictions on new data including transformations, imputation results, and model parameter estimates and equations.

Examples…

SAS Discharge Disposition and Length of Stay Modeling for Hospitals

- **Length of stay:** Survival modeling to predict 'target' discharge date up to 2 days prior to discharge for patients who end up going home with care or without care.

- **Discharge disposition**: Predict discharge disposition 2 days prior to patient discharge for those patients who will go home with home care and those who will go home without homecare.

- Data: Use daily in-hospital data from admissions, OR, DI, pharmacy, lab tests, etc. to score patients daily.

Predicting End Stage Renal Failure
Survival modeling to predict probability of developing End Stage Renal Disease given patient attributes and kidney function measures.

What makes these predictive models?

- Same algorithms as employed in inferential statistics, but different methodology and modeling ***purpose—to score individuals in near real time and use results for rapid and preemptive decision making***.

The North Carolina Birth Records Data

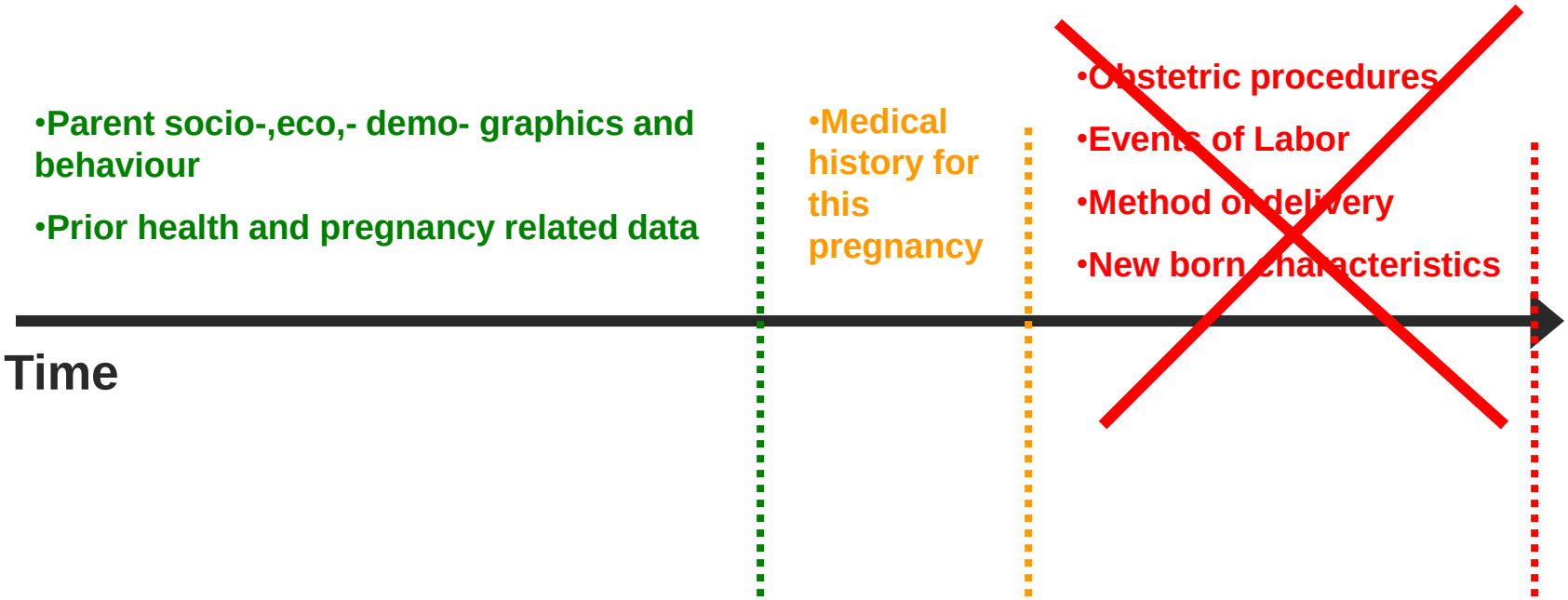
- North Carolina Birth Records from North Carolina Center for Health Statistics: 122,550 from 2000, and 120,300 from 2001.
- 7.2% low birth weight births (< 2500 grams) excluding multiple births.
- Data contains information on parents ethnicity, age, education level and marital status.
- Data contains information on mothers health condition and reproductive history.
- 45 potential predictor variables for modeling.

Scenario: Early Warning System for Birth Weight

PREDICTORS

- **Parent socio-,eco-, demo- graphics, health and behaviour**
  - **Age, edu, race, medical conditions, smoking, drinking etc.**
- **Prior pregnancy related data**
  - **# pregnancies, last outcome, prior pregnancies etc.**
- **Medical History for pregnancy**
  - **Hypertension during pregnancy, eclampsia, incompetent cervix, etc.**
- **Obstetric procedures**
  - **Amniocentesis, ultrasound, etc.**
- **Events of Labor**
  - **Breech, fetal distress etc.**
- **Method of delivery**
  - **Vaginal, c-section etc.**
- **New born characteristics**
  - **congenital anomalies (spinabifida, heart), APGAR score, anemia**

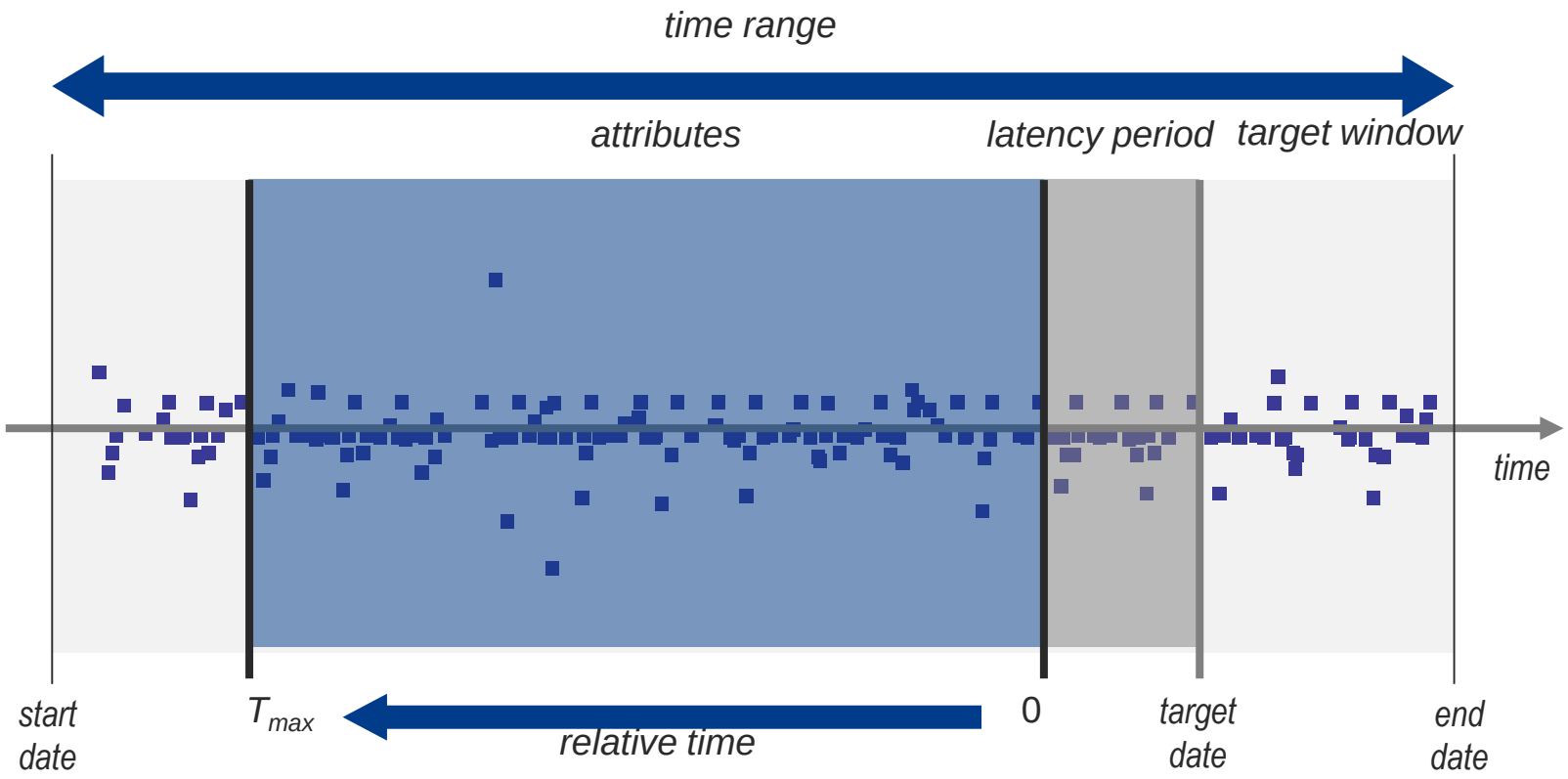
Beware of Temporal Infidelity…..



Data & Variable Preparation

THE POWER TO KNOW®

Preparing the Modeling Data



Oversample Rare Events

```
proc sort data=bwt00;
    by lbwt;
run;

proc surveyselect data=bwt00
            samprate=(.075,1) out=OSbwt00 seed=5;
    strata lbwt;
run;

proc freq data=OSbwt00;
    tables lbwt;
run;
```

lbwt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	8260	49.63	8260	49.63
1	8383	50.37	16643	100.00

- SURVEYSELECT is used to sample 7.5% of non-events and 100% of events.
- Data must be sorted by the target prior to oversampling.

Create Missing Indicators

```
data bwt00;
set bwt00;
  array vars{*} fage mage feduc meduc totalp bdead terms
    loutcome prenatal marital children cignum drinknum anemia
    cardiac aclung diabetes herpes hydram hemoglob hyperch
    hyperpr eclamp cervix pinfant preterm renal rhsen uterine
    amnio ultra YrsLastLiveBirth YrsLastFetalDeath drinker smoker marital;

  array mvars{*} M_fage M_mage M_feduc M_meduc M_totalp M_bdead M_terms
    M_loutcome M_prenatal M_marital M_children M_cignum M_drinknum M_anemia
    M_cardiac M_aclung M_diabetes M_herpes  M_hydram M_hemoglob M_hyperch
    M_hyperpr M_eclamp M_cervix M_pinfant M_preterm M_renal M_rhsen M_uterine
    M_amnio M_ultra M_YrsLastLiveBirth M_YrsLastFetalDeath M_drinker M_smoker M_marital;
     do i=1 to dim(vars);
      mvars{i}=(vars{i}=.);
   end;
run;
```

- Create missing indicators to capture associations between missingness and the target in development data.
- The process is repeated for Test data.
- This step is unnecessary for Decision Trees as they accommodate missing values directly.

Partition Data for Empirical Validation

```
proc surveyselect
                    data=OSbwt00
                    samprate=.6667
                    out=dev00
                    seed=44444
                    outall;
                    strata lbwt;
run;
```

```
data train valid;
   set dev00;
   if selected then output train;
   else output valid;
 run;
```

- SURVEYSELECT is used to partition data into Training (67%) and Validation (33%) sets.
- The OUTALL option provides one dataset with a variable, SELECTED that indicates dataset membership.
- For class targets, stratification on the target, LBWT ensures equal representation of low birth weight cases in training and validation sets.
- Since HPSPLIT and HPGENSELECT procedures do not accept separate train and validate sets, the dataset output from SURVEYSELECT will be used before physically splitting into train and validation data.

Impute Missing Values

```
proc stdize data=train reponly method=median out=trainI outstat=med;
   var &numvars;
run;

proc stdize data=valid out=validI reponly method=in(med);
    var &numvars;
run;

proc stdize data=test01 out=testI reponly method=in(med);
    var &numvars;
run;
```

- STDIZE will replace missing values (REPONLY) and is applied to the Training data.
- The OUTSTAT option saves a dataset to be used to insert results (score) into Validation and Test sets.
- The METHOD=IN (MED) uses the imputation information from the training data to score the Validation and Test data.
- Imputation is unnecessary for Decision Trees as they accommodate missing values directly.

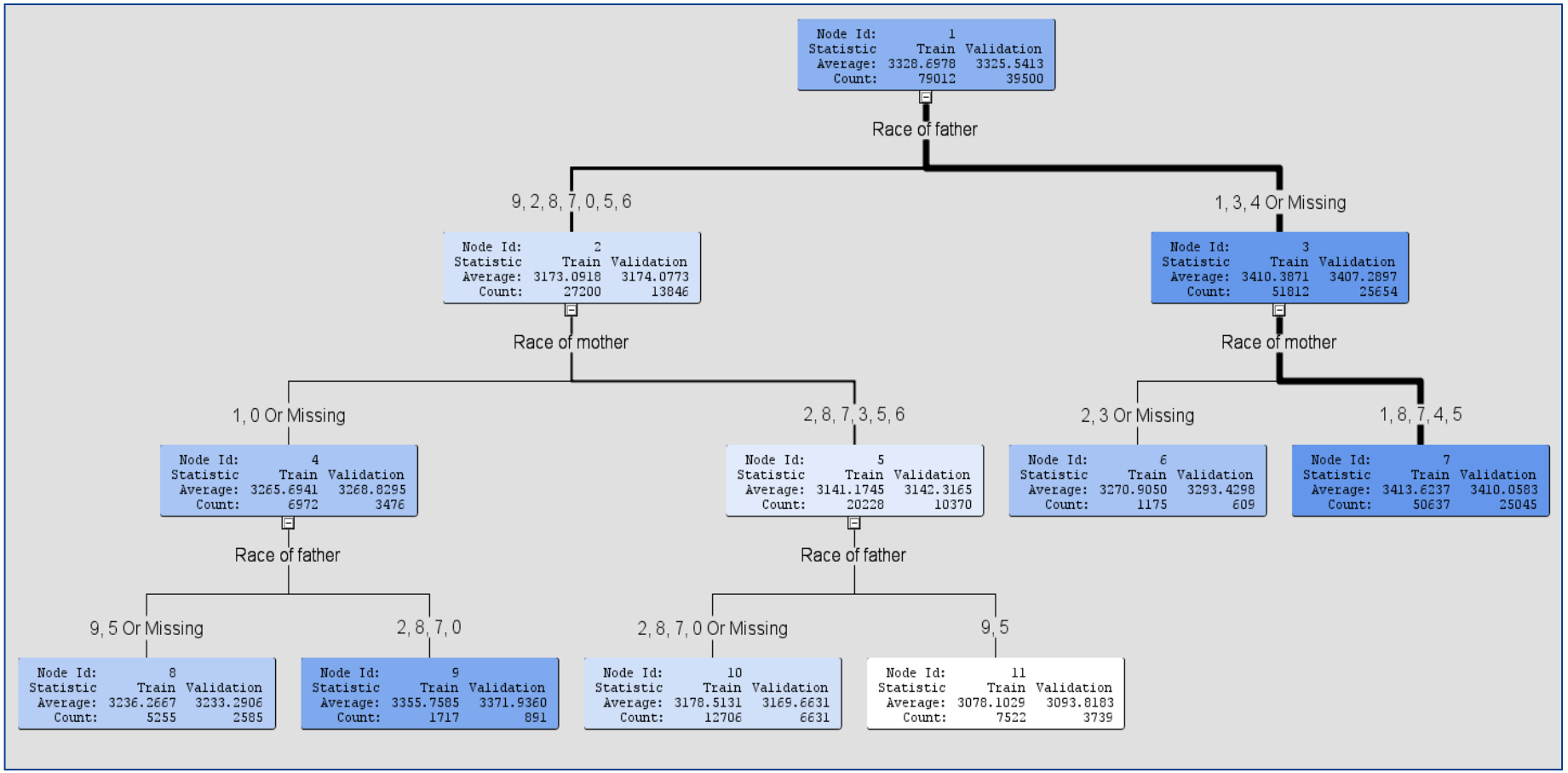
Cluster Variables to Reduce Dimensions

Cluster 2	CHILDREN	0.8025	0.0916	0.2174	Number of children now living
	LOUTCOME	0.7907	0.0801	0.2275	Outcome of last delivery
	TOTALP	0.7398	0.4152	0.4449	Total pregnancies (including this one)
	M_YrsLastLiveBirth	0.8245	0.0782	0.1904	
Cluster 3	FEDUC	0.8554	0.1557	0.1713	Education of father (years)
	MEDUC	0.8554	0.1737	0.1751	Education of mother (years)
Cluster 4	M_terms	0.9329	0.0830	0.0732	
	M_totalp	0.9329	0.2496	0.0894	
Cluster 5	M_cignum	0.9537	0.0707	0.0499	
	M_drinknum	0.9537	0.0859	0.0507	
Cluster 6	M_fage	0.9257	0.0381	0.0772	
	M_feduc	0.9244	0.0379	0.0786	
	MARITAL	0.5320	0.1420	0.5455	Marital status
Cluster 7	CIGNUM	0.8454	0.2511	0.2065	Average # of cigarettes daily
	smoker	0.8454	0.0268	0.1589	
Cluster 8	DRINKNUM	0.6777	0.0007	0.3225	Average # of alcoholic drinks per week
	drinker	0.6777	0.2236	0.4151	

- Cluster variables on training data to reduce collinearity prior to modeling. E.g. PROC VARCLUS.

```
proc varclus data=train maxeigen=.7 short hi;
 var  &IntervalandFlagVars;
 run;
```

Collapse Categorical Variables to Reduce Dimensions



- Variables RACEMOM and RACEDAD contain 9 and 10 levels respectively.
- Use a Decision Tree model to optimally collapse many possible combinations of these attributes to a single 6-level variable using training data.
- This step is unnecessary if you are using a decision tree as a predictive model.

Binary Target Modeling

New Modeling Routines in SAS/STAT: Decision Trees using HPSPLIT

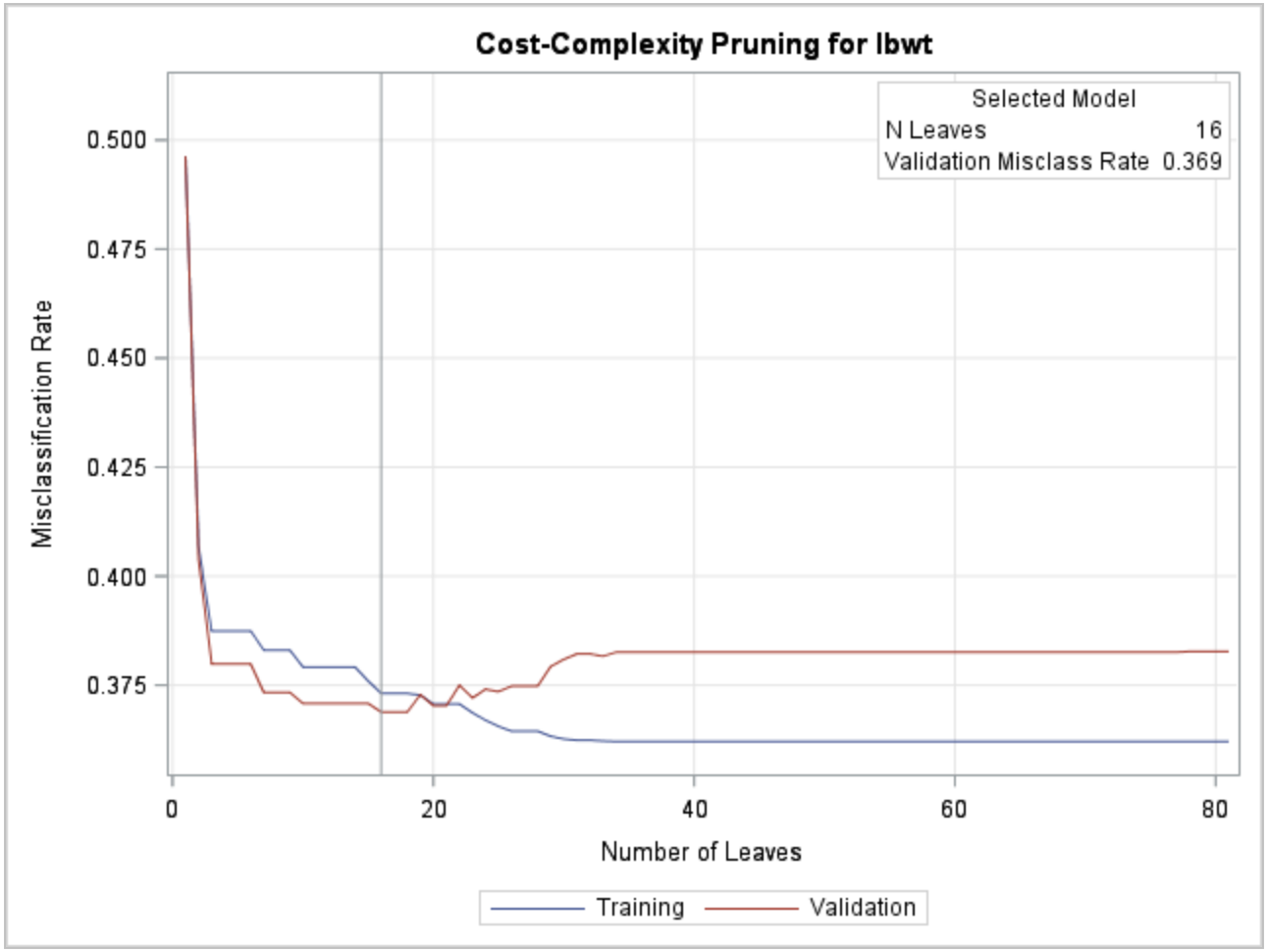
- In SAS 9.4 some of the high performance procedures used in Enterprise Miner software for data mining are now available in SAS/STAT (now called version 14.1).
- Procedures support parallel processing and are designed to run in a distributed computing environment (across multiple servers for high speed computing).
- Procedures are available to run on a single machine or server. They are now shipped with SAS/STAT at no additional cost.
- In this section we will feature HPSPLIT for decision trees using a binary target and in a later section, HPGENSELECT for generalized linear models and mixture distributions using a continuous target.

Build Decision Trees using HPSPLIT

```
proc hpsplit data=dev00 plots=zoomedtree(nodes=("Q","C"))
        minleafsize=100 nsurrogates=2 assignmissing=popular
        nodes maxbranch=2 mincatsize=50 maxdepth=12;
    class &classvarstree lbwt;
    model lbwt(event='1')= &numvarstree &classvarstree;
    partition roleVar=Selected(train='1' validate='0');
    grow gini;
    prune costcomplexity;
    code file='C:\data\EDU\TALKS\UG16\lbwtTREE.sas';
run;
```

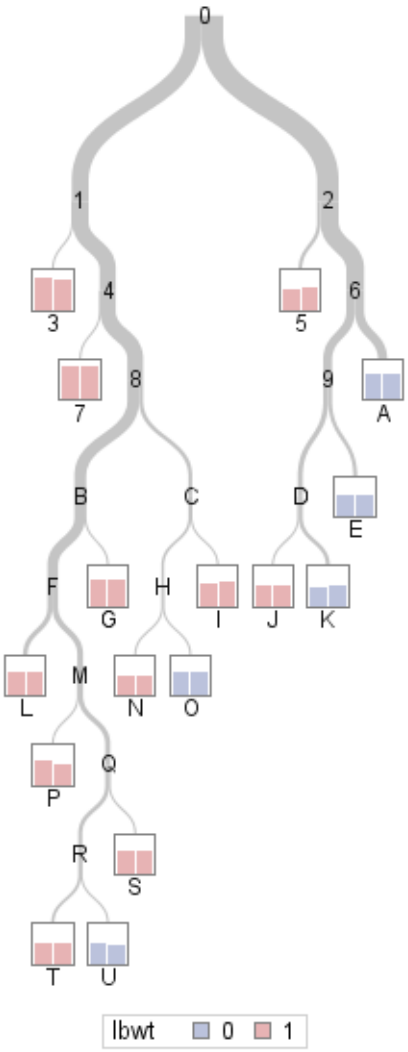
- The program fits a CART-like decision tree to low birth weight data: with surrogates, GINI splitting criterion, Cost-complexity pruning (Breiman et al.), and data splitting (PARTITION) from a single development dataset using a flag variable (SELECTED) that indicates train/validate membership.
- CHAID-like and machine learning-like trees are also possible.
- Tree plots are subset using a ZOOMEDTREE option. Node details are printed using a NODES options.
- Minimum leaf and class variable sizes, maximum branches, and maximum depth are set using MINLEAFSIZE, MINCATSIZE, MAXBRANCH, and MAXDEPTH options.
- Missing values are assigned to the branch with largest sample size (ASSIGNMISSING=POPULAR).
- Data Step scoring code is saved to a file using the CODE statement, available in PROCs: GENMOD, GLIMMIX, GLM, GLMSELECT, LOGISTIC, MIXED, REG, HPSPLIT, HPGENSELECT and others…

Cost-Complexity Pruning with the PRUNE Statement



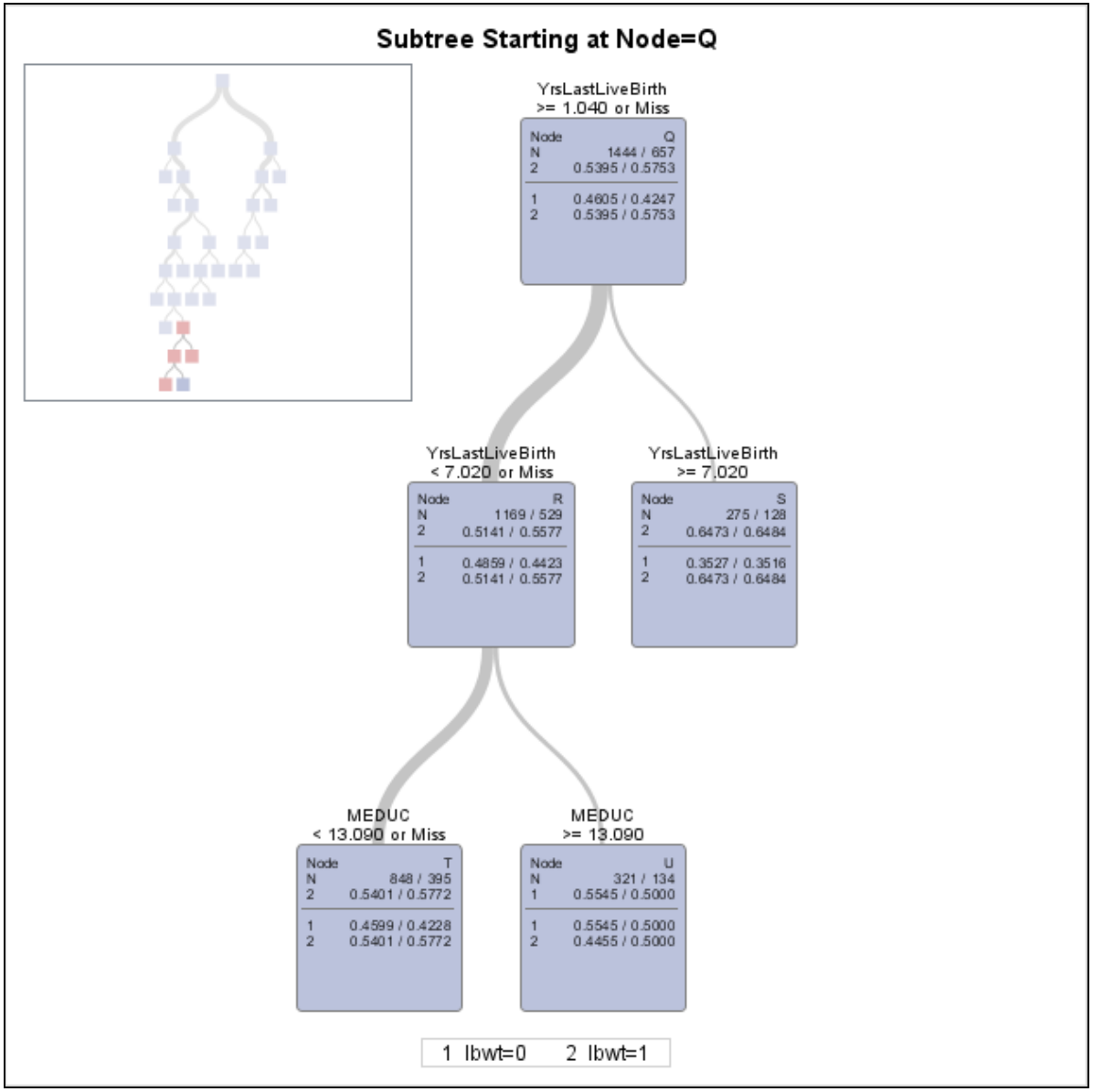
- Though the full tree has over 80 leaves, Cost-Complexity Pruning on misclassification rate yields a 16 leaf tree that minimizes misclassification rate.

The Final Tree



- Nodes and leaves are identified with numbers and letters.
- The width of the curves are proportional to amount of data passing through each part of the tree.
- Red indicates low bwt classification while blue indicates normal weight.

Plot Tree Sections Using the ZOOMEDTREE option



- Classifications using a cutoff of 0.5 are given at the top of each node (e.g. Node T is classified as low birth weight while Node U is classified as normal weight).
- Samples sizes for Train/Validate data, as well as proportions of target 1 in each are shown.
- Red indicates low bwt classification while blue indicates normal weight.

Display Leaf Rules using the NODES Option

Node Information									
		Training Data				Validation Data			
ID	Path	Count	0		1		Count	0	1
3	Root Node	11096	0.4963		0.5037		5547	0.4963	0.5037
	RACEDAD = 2,3,8,9	4858	0.3895		0.6105		2402	0.3847	0.6153
	PRENATAL < 0.09	192	0.1250		0.8750	*	104	0.1827	0.8173
5	Root Node	11096	0.4963		0.5037		5547	0.4963	0.5037
	RACEDAD = 0,1,4,5,6,7 or Missing	6238	0.5795		0.4205		3145	0.5816	0.4184
	smoker = 1	1071	0.4006		0.5994	*	573	0.3857	0.6143
7	Root Node	11096	0.4963		0.5037		5547	0.4963	0.5037
	RACEDAD = 2,3,8,9	4858	0.3895		0.6105		2402	0.3847	0.6153
	PRENATAL >= 0.09 or Missing	4666	0.4003		0.5997		2298	0.3938	0.6062
	PRETERM = 1	147	0.1088		0.8912	*	75	0.1200	0.8800
A	Root Node	11096	0.4963		0.5037		5547	0.4963	0.5037
	RACEDAD = 0,1,4,5,6,7 or Missing	6238	0.5795		0.4205		3145	0.5816	0.4184
	smoker = 0 or Missing	5167	0.6166		0.3834		2572	0.6252	0.3748

… etc etc.

- Rules details for all leaves are reported. Asterisk indicates selected target level.

Explore Variable Importance

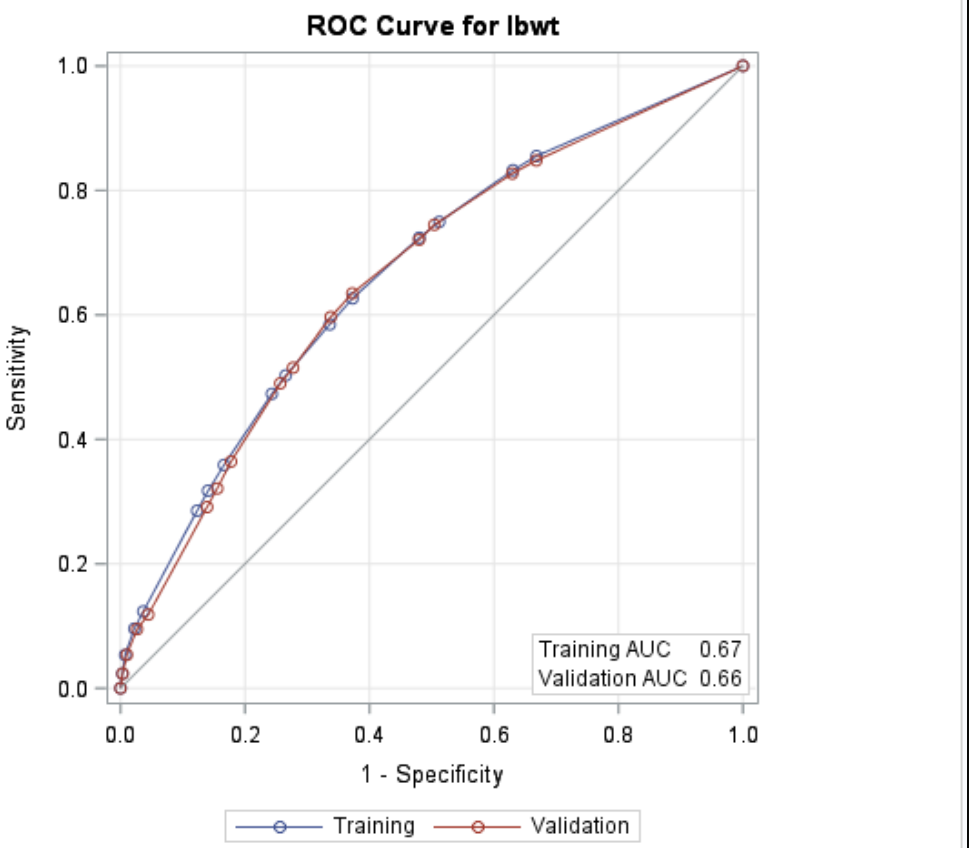
Variable Importance								
Variable	Variable Label	Training		Validation		Relative Ratio	Count	Times Used as a Surrogate
		Relative	Importance	Relative	Importance			
RACEMOM	Race of mother	1.0000	15.7420	1.0000	10.9894	1.0000	1	4
RACEDAD	Race of father	0.9258	14.5744	0.9661	10.6169	1.0435	1	2
MARITAL	Marital status	0.7789	12.2611	0.8156	8.9634	1.0472	0	1
CIGNUM	Average # of cigarettes daily	0.6946	10.9339	0.7792	8.5631	1.1219	3	1
smoker		0.6556	10.3209	0.7416	8.1502	1.1312	0	3
drinker		0.5687	8.9527	0.6257	6.8757	1.1001	0	2
CHILDREN	Number of children now living	0.5918	9.3157	0.4677	5.1398	0.7903	3	0
TOTALP	Total pregnancies (including this one)	0.5625	8.8553	0.4445	4.8843	0.7901	0	3
MAGE	Age of mother	0.5500	8.6587	0.4154	4.5652	0.7552	0	7
PRETERM	Prev. preterm/small	0.3226	5.0788	0.3099	3.4058	0.9606	1	0
DRINKNUM	Average # of alcoholic drinks per week	0.3172	4.9936	0.3047	3.3486	0.9606	0	1
PRENATAL	Month of preg. prenatal care began	0.3598	5.6632	0.2945	3.2365	0.8186	2	0
TERMS	Number of other terminations	0.2155	3.3926	0.2727	2.9972	1.2655	0	2
MEDUC	Education of mother (years)	0.2124	3.3440	0.1767	1.9420	0.8319	1	1
FEDUC	Education of father (years)	0.1879	2.9586	0.1594	1.7513	0.8479	1	0
RENAL	Renal disease	0.1151	1.8113	0.1255	1.3790	1.0906	0	1
YrsLastLiveBirth		0.2375	3.7390	0.1070	1.1761	0.4506	2	0
BDEAD	Number previous live births now dead	0.1606	2.5284	0.0963	1.0583	0.5996	0	1

- Race of mother and father as well as marital status and smoking behavior are the top most important variables.

Model Assessments

Fit Statistics for Selected Tree

	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Training	16	0.2278	0.3731	0.6264	0.6274	0.9331	0.4556	5055.7	0.6674
Validation	16	0.2291	0.3688	0.6346	0.6277	0.9355	0.4568	2541.6	0.6627



- Training and Validation assessment measures and overlayed ROC curves are output.

Score Data using HPSPLIT Scoring Code

```
data scoredtreeval(keep=p_tree p_1adjtree idnum);
    set DEV00;
    %include 'C:\data\EDU\TALKS\HUG2015\lbwtTREE.sas'/source2;
    P_1ADJtree=(P_lbwt1*(1-&RHO1)*&PRIOR1)/
        ((1-P_LBWT1)*&RHO1*(1-&PRIOR1)+(P_LBWT1*(1-&RHO1)*&PRIOR1));
    if selected=0;
    p_Tree=p_lbwt1;
run;
```

Etc

```
data scoredtreetest(keep=p_tree p_1adjtree idnum);
    set test01;
    %include 'C:\data\EDU\TALKS\HUG2015\lbwtTREE.sas'/source2;
    P_1ADJtree=(P_lbwt1*(1-&RHO1)*&PRIOR1)/
        ((1-P_LBWT1)*&RHO1*(1-&PRIOR1)+(P_LBWT1*(1-&RHO1)*&PRIOR1));
    p_Tree=p_lbwt1;
run;
```

Etc

- Validation and Test data are scored using Decision Tree Scoring code.
- Probabilities can be adjusted for oversampling (P_1ADJtree) if desired though this is not required for ROC curve assessments.
- Validation and Test scores from the tree model are match-merged back to the corresponding imputed sets to be used for regression (not shown here).

Select Regression Models and Score

```
title "Backward Early Warning Regression";
proc logistic data=trainI;
   class tree_race(param=ref ref='8');
   model lbwt(event='1')=&numvars2 tree_race/ selection = backward slstay=.01;
   score data=allval out=sco_validate(rename=(p_1=p_early))priorevent=.072;
   score data=alltest out=sco_test(rename=(p_1=p_early))priorevent=.072;
run;
title "Backward Regression";
proc logistic data=trainI;
    class tree_race(param=ref ref='8');
    model lbwt(event='1')=&numvars tree_race/selection = backward slstay=.01;
    score data=sco_validate out=sco_validate(rename=(p_1=p_all)) priorevent=.072;
    score data=sco_test out=sco_test(rename=(p_1=p_all)) priorevent=.072;
run;
title "Backward Interactions Regression";
proc logistic data=trainI;
    class tree_race(param=ref ref='8');
    model lbwt(event='1')= HYPERCH HYPERPR CERVIX BDEAD CIGNUM ECLAMP HEMOGLOB HYDRAM MEDUC
        PINFANT PRENATAL PRETERM RHSEN TOTALP UTERINE drinker M_YrsLastLiveBirth
        M_smoker MARITAL smoker Tree_Race HYPERCH|HYPERPR|CERVIX|BDEAD|CIGNUM|ECLAMP|
        HEMOGLOB|HYDRAM MEDUC|PINFANT|PRENATAL|PRETERM|RHSEN|TOTALP|UTERINE|drinker|M_YrsLastLiveBirth|
        M_smoker|MARITAL|smoker|Tree_Race @2/ selection = forward slentry=.01 include=21 ;
    score data=sco_validate out=sco_validate(rename=(p_1=p_AllInt))priorevent=.072;
    score data=sco_test out=sco_test(rename=(p_1=p_AllInt))priorevent=.072;
run;
```

- The SCORE statements allows for scoring of new data (Validation and Test) and adjusts oversampled data back to the population prior (PRIOREVENT=0.072).

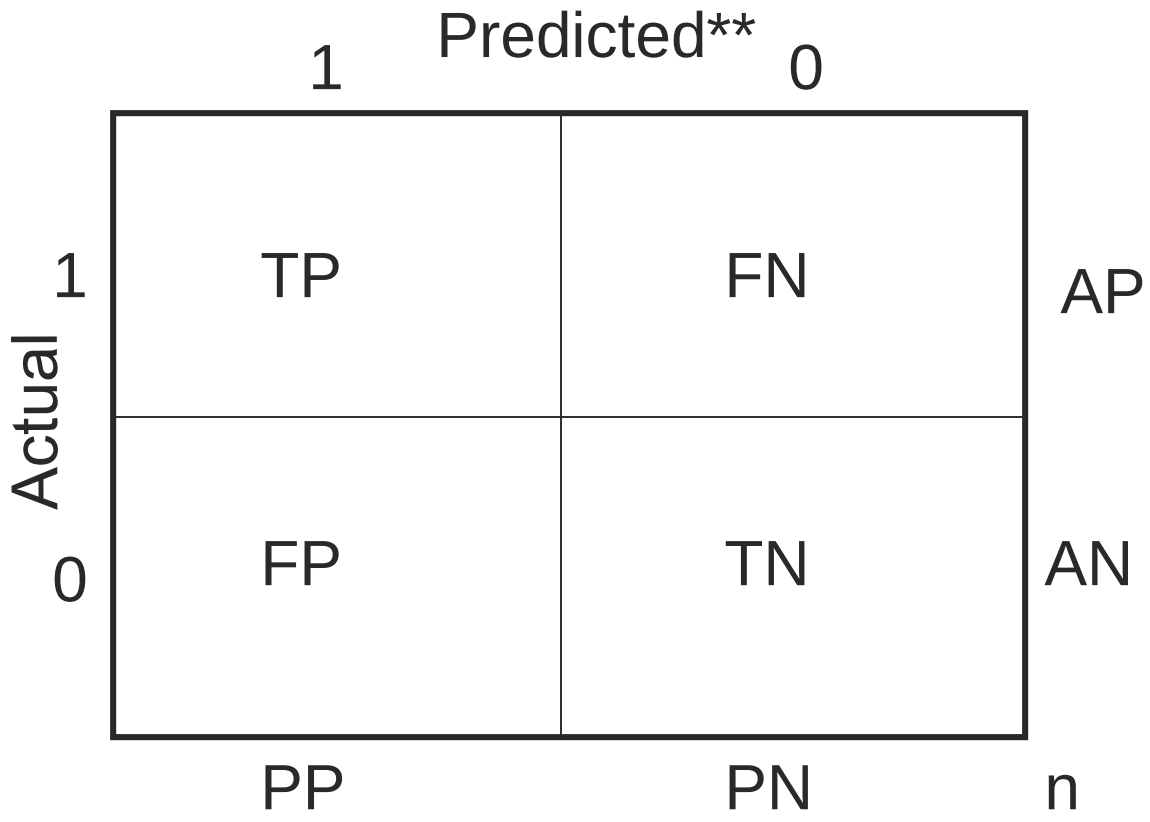
- The ALLVAL and ALLTEST sets containing decision tree predictions are supplied in the first regression run. The same datasets are re-scored (SCO_VALIDATE, SCO_TEST), and prediction variables renamed, so that predictions for all four models are in the same set for comparisons.

Early Warning Regression Output, for example..

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
HYPERCH	2.930	1.981	4.334
CIGNUM	1.028	1.014	1.042
FEDUC	0.969	0.948	0.991
MAGE	1.011	1.003	1.019
MEDUC	0.969	0.949	0.990
PINFANT	0.266	0.129	0.549
PRENATAL	0.891	0.865	0.918
PRETERM	7.270	5.091	10.382
RENAL	3.008	1.378	6.565
TOTALP	1.089	1.051	1.129
M_YrsLastLiveBirth	2.042	1.842	2.262
MARITAL	1.214	1.089	1.354
smoker	1.522	1.274	1.818
Tree_Race 6 vs 8	1.356	0.829	2.218
Tree_Race 7 vs 8	0.806	0.682	0.953
Tree_Race 9 vs 8	0.930	0.684	1.264
Tree_Race 10 vs 8	1.849	1.550	2.204
Tree_Race 11 vs 8	2.114	1.758	2.543
Tree_Race 99 vs 8	1.242	0.825	1.872

- In general, predictive models fit to larger datasets tend to have more parameters than more theoretically informed explanatory models in health.
- Odds ratios for previous premature babies (PRETERM), renal disease (RENAL), and chronic hypertension (HYPERCH) and are particularly large.
- The collapsed version of mother and father race from an initial decision tree (TREE_RACE) appears in the model.

Model Assessments for Binary Targets



Accuracy = (TP+TN)/n

Sensitivity = TP/AP

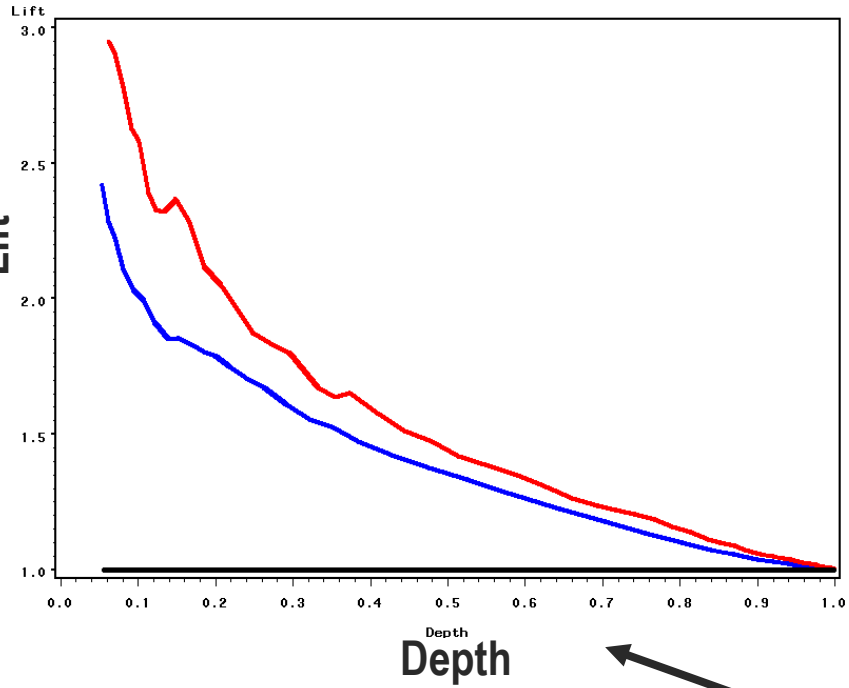
Specificity = TN/AN

Lift = (TP/PP)/π_1

** - **Where Predicted 1=(Pred Prob > Cutoff)**

Assessment Charts for Binary Targets

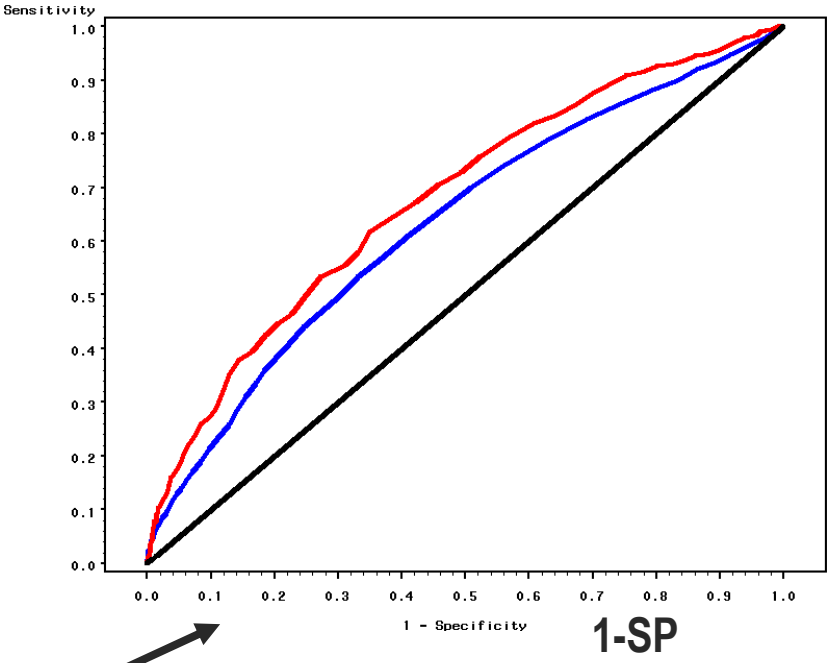
Lift Charts



Lift

Depth

ROC Charts



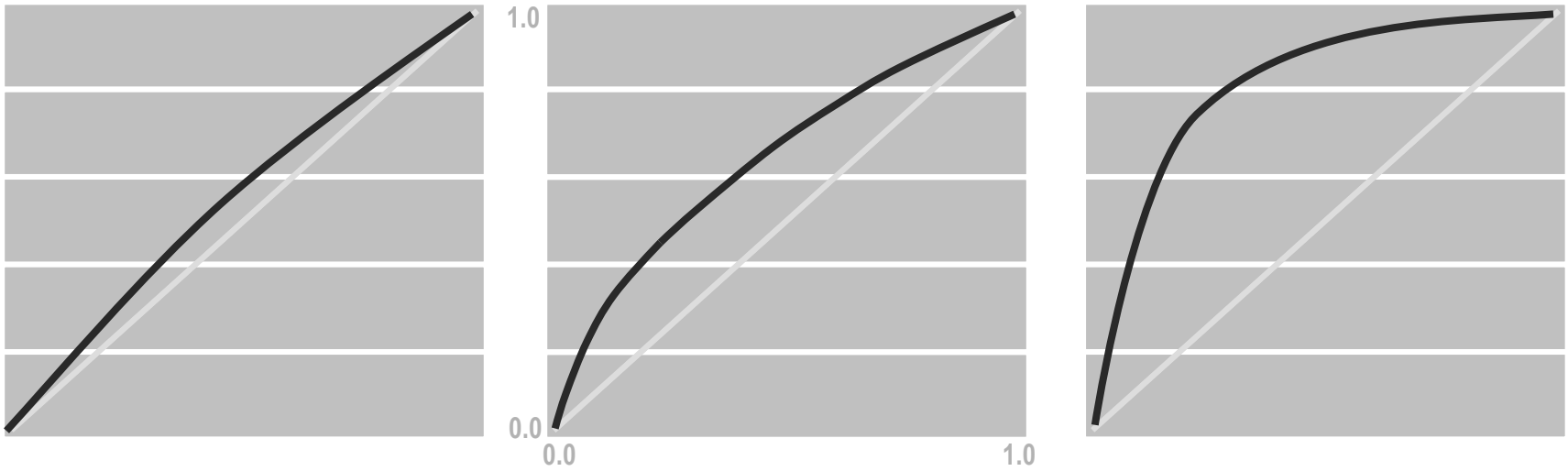
SE

1-SP

Explore measures across a range of cutoffs

TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN
FP	TN	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN

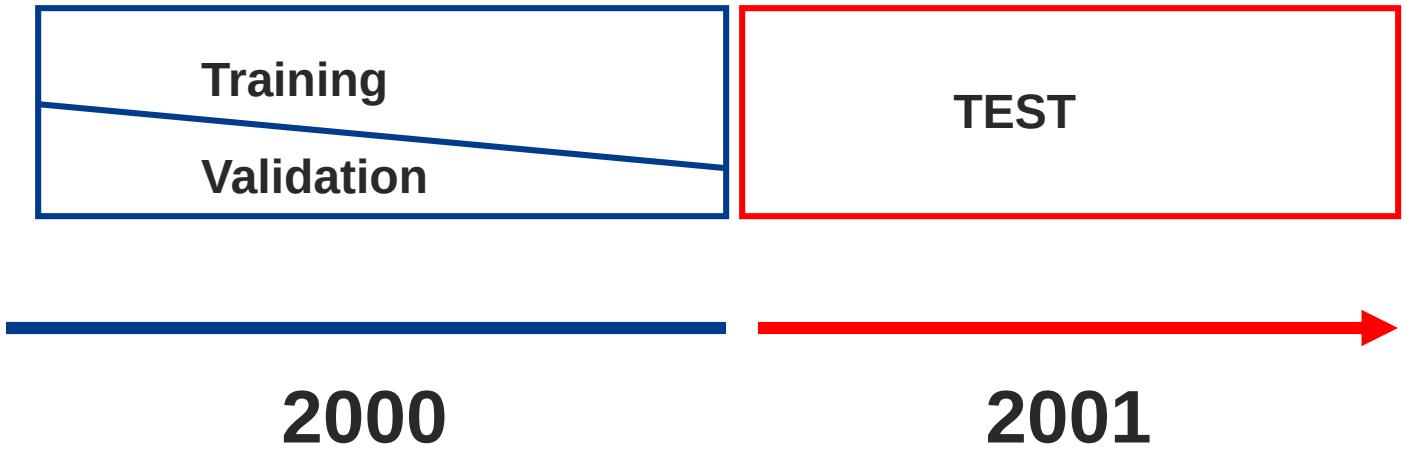
Receiver Operator Curves



weak model **strong model**

- A measure of a model's predictive performance, or model's ability to discriminate between target class levels. Areas under the curve range from 0.5 to 1.0.
- A concordance statistic: for every pair of observations with different outcomes (LBWT=1, LBWT=0) AuROC measures the probability that the ordering of the predicted probabilities agrees with the ordering of the actual target values.
- …Or the probability that a low birth weight baby (LBWT=1) has a higher predicted probability of low birth weight than a normal birth weight baby (LBWT=0).

Predict the Future with Data Splitting



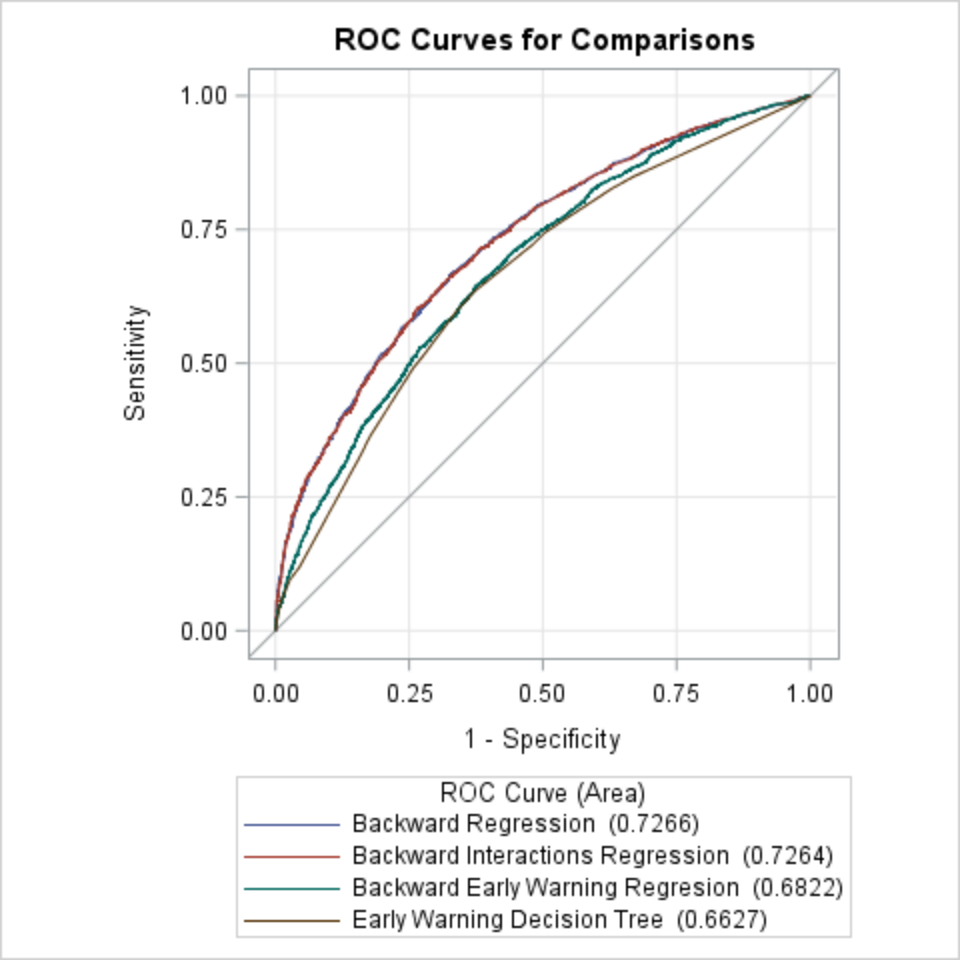
- Models are fit to Training Data, compared and selected on Validation and tested on a future Test set.

Assess Models using ROC Curves

```
title "Comparing Four Models on Validation Data";
ods html;
ods graphics on;
proc logistic data=sco_validate;
   model lbwt(event='1')=p_all p_allint p_early p_tree / nofit;
   roc "Backward Regression" p_all;
   roc "Backward Interactions Regression" p_allInt;
   roc "Backward Early Warning Regresion" p_early;
   roc "Early Warning Decision Tree" p_Tree;
   roccontrast "Comparing the Four Models: Validation Data "/estimate=allpairs;
run;
```

- The dataset with all four predictions (SCO_VALIDATE) is supplied to PROC LOGISTIC.
- The ROCCONTRAST statements provides hypothesis tests for differences between ROC curves, for model results specified in the three ROC statements.
- To generate ROC contrasts, all terms used in the ROC statements must be placed on the model statement. The NOFIT option suppresses the fitting of the specified model.
- Because of the presence of the ROC and ROCCONTRAST statements, ROC plots are generated when ODS GRAPHICS are enabled.
- The identical process is repeated with the scored Test set, SCO_TEST. Can the model predict the future?

Compare ROC Curves on Validation Data



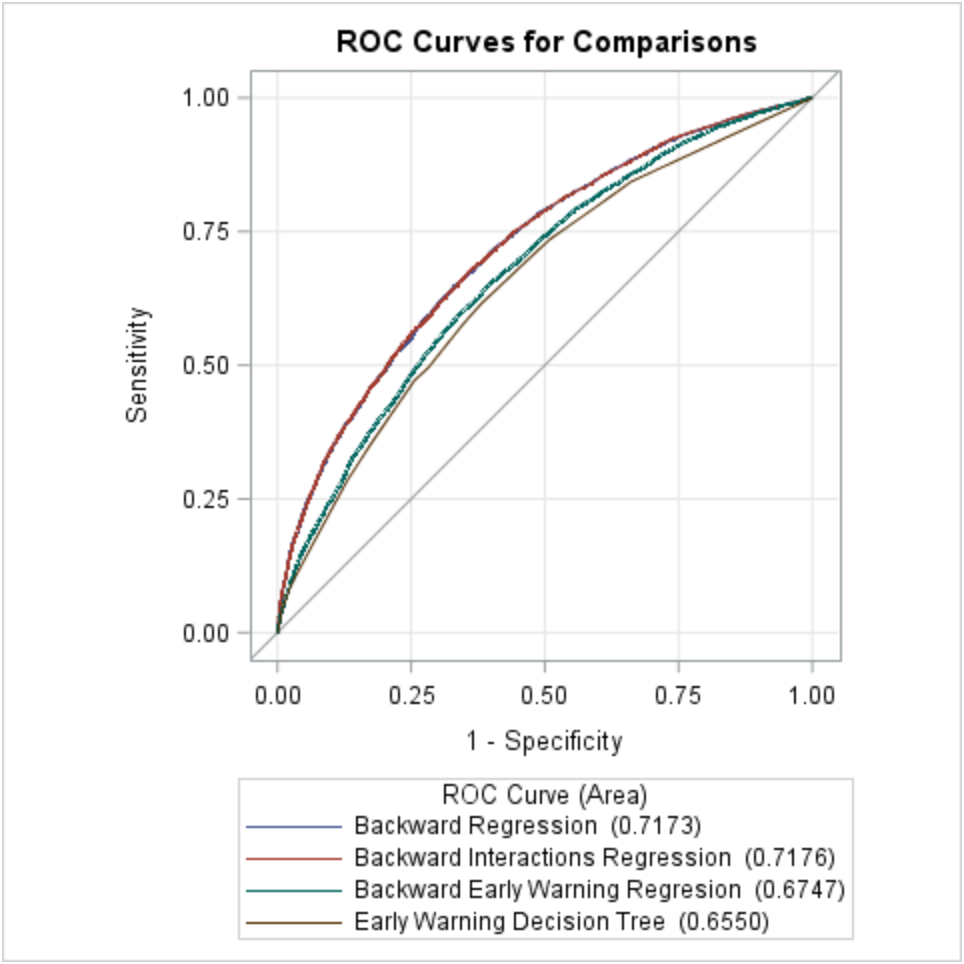
Compare AuROC Curves on Validation Data

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Backward Regression	0.7266	0.00670	0.7135	0.7397	0.4532	0.4540	0.2266
Backward Interactions Regression	0.7264	0.00670	0.7133	0.7396	0.4529	0.4536	0.2265
Backward Early Warning Regresion	0.6822	0.00708	0.6684	0.6961	0.3645	0.3645	0.1823
Early Warning Decision Tree	0.6627	0.00719	0.6486	0.6768	0.3253	0.3644	0.1627

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Comparing the Four Models: Validation Data	3	147.7898	<.0001

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
Backward Regression - Backward Interactions Regression	0.000164	0.00107	-0.00193	0.00226	0.0234	0.8784
Backward Regression - Backward Early Warning Regresion	0.0443	0.00422	0.0361	0.0526	110.1907	<.0001
Backward Regression - Early Warning Decision Tree	0.0639	0.00565	0.0528	0.0750	127.8547	<.0001
Backward Interactions Regression - Backward Early Warning Regresion	0.0442	0.00433	0.0357	0.0527	103.8641	<.0001
Backward Interactions Regression - Early Warning Decision Tree	0.0638	0.00561	0.0528	0.0747	129.2086	<.0001
Backward Early Warning Regresion - Early Warning Decision Tree	0.0196	0.00444	0.0109	0.0283	19.4494	<.0001

Compare ROC Curves on Test Data



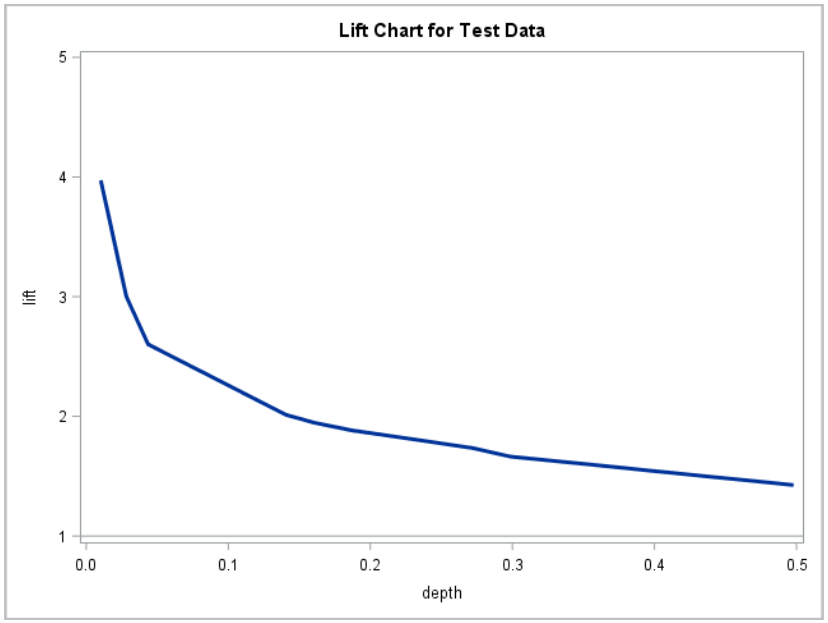
Compare AuROC Curves on Test Data

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Backward Regression	0.7173	0.00392	0.7097	0.7250	0.4347	0.4353	0.2173
Backward Interactions Regression	0.7176	0.00392	0.7100	0.7253	0.4353	0.4359	0.2176
Backward Early Warning Regresion	0.6747	0.00412	0.6666	0.6828	0.3494	0.3494	0.1747
Early Warning Decision Tree	0.6550	0.00416	0.6468	0.6631	0.3100	0.3477	0.1550

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Comparing the Four Models: Test Data	3	401.7402	<.0001

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
Backward Regression - Backward Interactions Regression	-0.00029	0.000570	-0.00141	0.000828	0.2584	0.6112
Backward Regression - Backward Early Warning Regresion	0.0426	0.00248	0.0378	0.0475	294.6697	<.0001
Backward Regression - Early Warning Decision Tree	0.0624	0.00338	0.0557	0.0690	341.1706	<.0001
Backward Interactions Regression - Backward Early Warning Regresion	0.0429	0.00253	0.0380	0.0479	287.3259	<.0001
Backward Interactions Regression - Early Warning Decision Tree	0.0627	0.00334	0.0561	0.0692	351.3665	<.0001
Backward Early Warning Regresion - Early Warning Decision Tree	0.0197	0.00269	0.0144	0.0250	53.8845	<.0001

Compare Lift Charts on Test Data

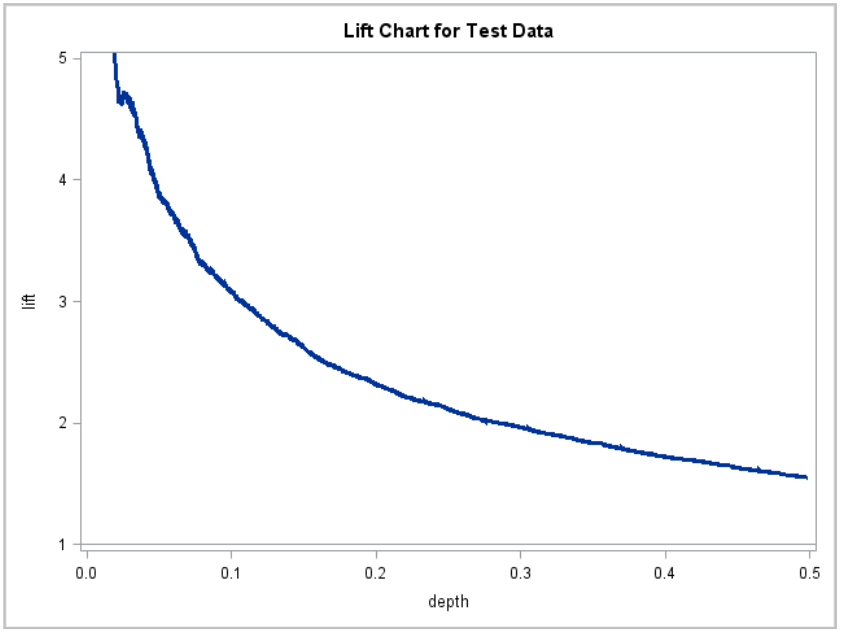


Early Warning Decision Tree

- Individuals in the top 5% most likely to have low birth weight babies are about **2.5** x more likely than average to have a lbwt baby.

All Effects Regression

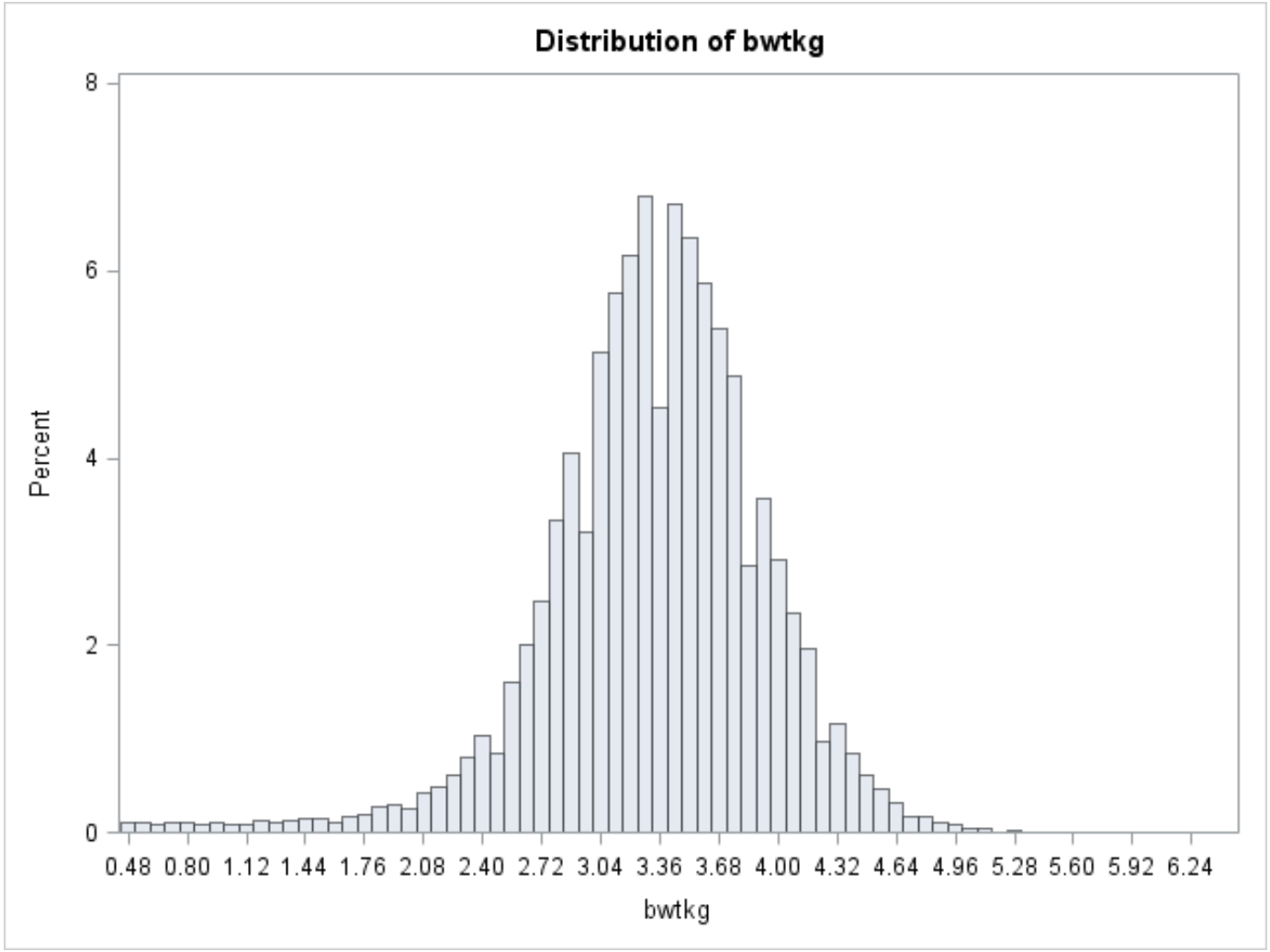
- Individuals in the top 5% most likely to have low birth weight babies are about **3.5** x more likely than average to have a lbwt baby.



Continuous Target Modeling

THE POWER TO KNOW.®

Continuous Target: Birth Weight



Build Regression Models with GLMSELECT

- GLMSELECT fits continuous target models (under GLM assumptions) and can process validation and test datasets, or perform cross validation for smaller datasets. It can also perform data partition using the PARTITION statement.
- GLMSELECT supports a class statement similar to PROC GLM but is designed for predictive modeling.
- Selection methods include Backward, Forward, Stepwise, LAR and LASSO.

Least Angle Regression

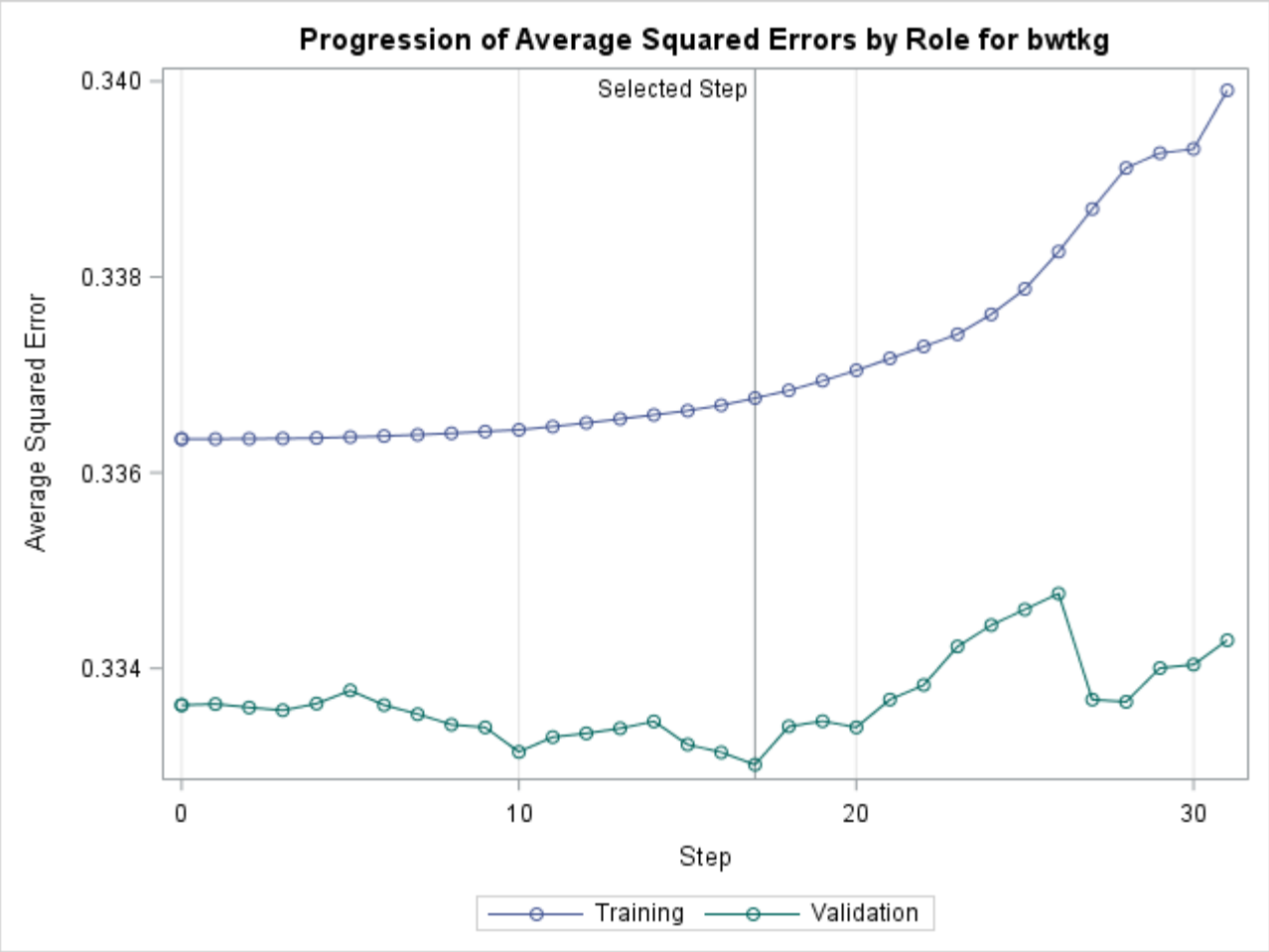
- Standardize inputs and response. All coefficients are zero.
- The predictor, X1 that is most correlated with the current residual (*makes the least angel with the residual*) is determined and a step is taken in the direction of this predictor (*X1 is added to model*). The length of this step (*the coefficient magnitude*) is chosen so that some other predictor, X2 and the current predicted response have the same correlation with the current residual (*equiangular*).
- At this point, the predicted response moves in the direction that is equiangular between (*equally correlated with*) these two predictors. Moving in this direction ensures that these two predictors continue to have a common correlation with the current residual.
- The predicted response moves in this direction until a third predictor, X3 has the same correlation with the current residual as the two predictors already in the model. A new direction is determined that is equiangular between these three predictors and the predicted response moves in this direction until a fourth predictor joins the set having the same correlation with the current residual.
- This process continues until all predictors are in the model.

Select Models with GLMSELECT

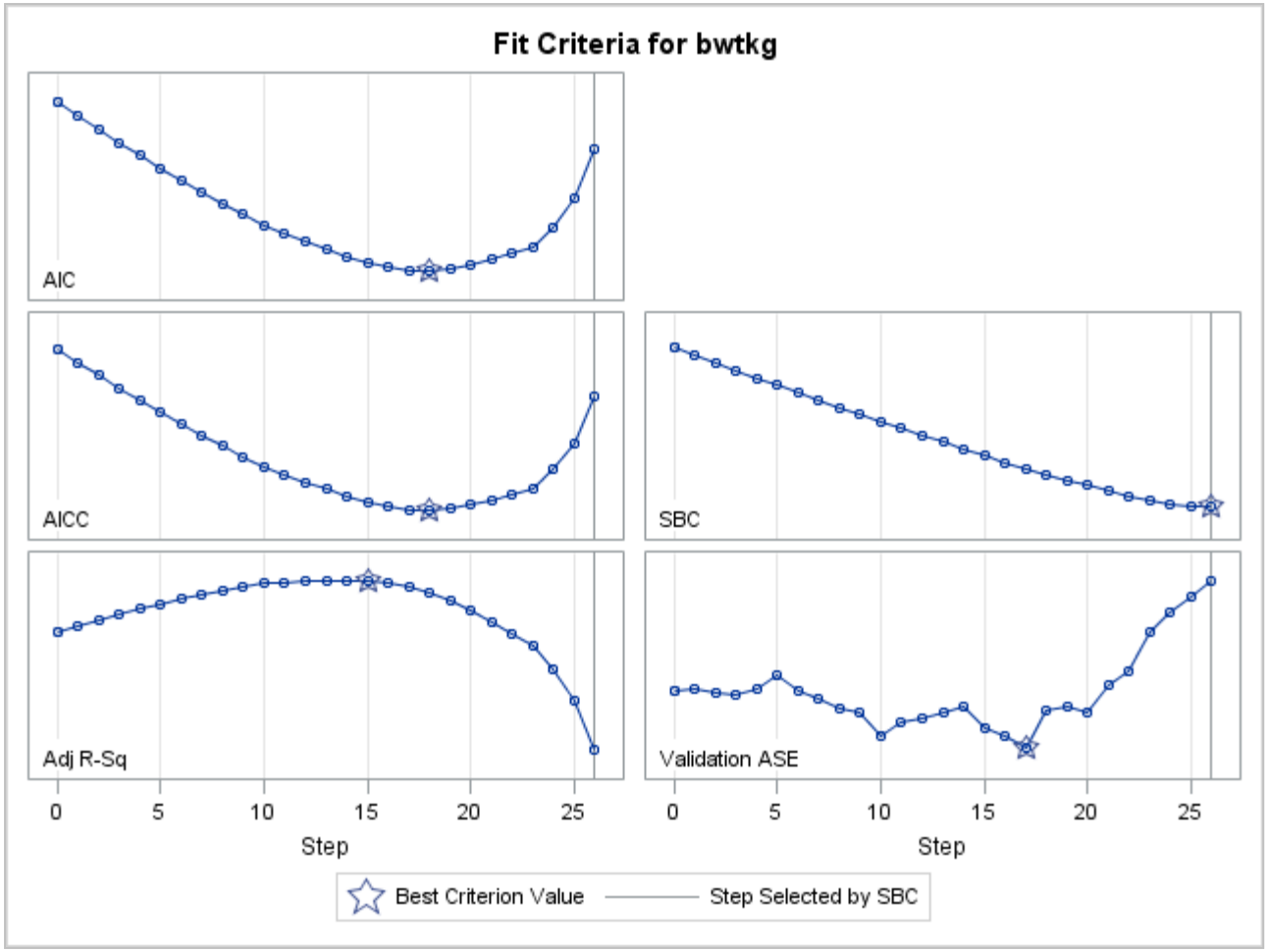
```
title1 'BACKWARD: select SL choose validate';
ods graphics on;
proc glmselect data=lbwt.train valdata=lbwt.valid /*testdata=test01*/
        plots(stepAxis=number)=ASEPlot;
    class tree_race;
    model bwtkg =  &numvars tree_race
        /selection=backward(choose = validate select = sl slstay=0.00001)
        showpvalues;
run;
title1 'BACKWARD: select SBC choose SBC';
 ods graphics on;
proc glmselect data=lbwt.train valdata=lbwt.valid
        plots=all;
    class tree_race;
    model bwtkg =  &numvars tree_race
        /selection=backward(choose = sbc select = sbc) showpvalues;
run;
title1 'LAR: Choose Validation ASE';
ods graphics on;
proc glmselect data=lbwt.train valdata=lbwt.valid
        plots=all;
    class tree_race;
    model bwtkg =  &numvars tree_race marital
         /selection=lar(choose = validate stop=none);
run;
```

- SELECTION=LAR requests Least Angle Regression.
- SELECT= determines the order in which effects enter or leave the model. Options include, for example: ADJRSQ, AIC, SBC, CP, CV, RSQUARE and SL. SL uses the traditional approach of significance level. SELECT is not available for LAR and LASSO.
- Models can be tuned with the CHOOSE= option to select the step in a selection routine using e.g. AIC, SBC, Mallow's CP, or validation data error. CHOOSE=VALIDATE selects that step that minimizes Validation data error.

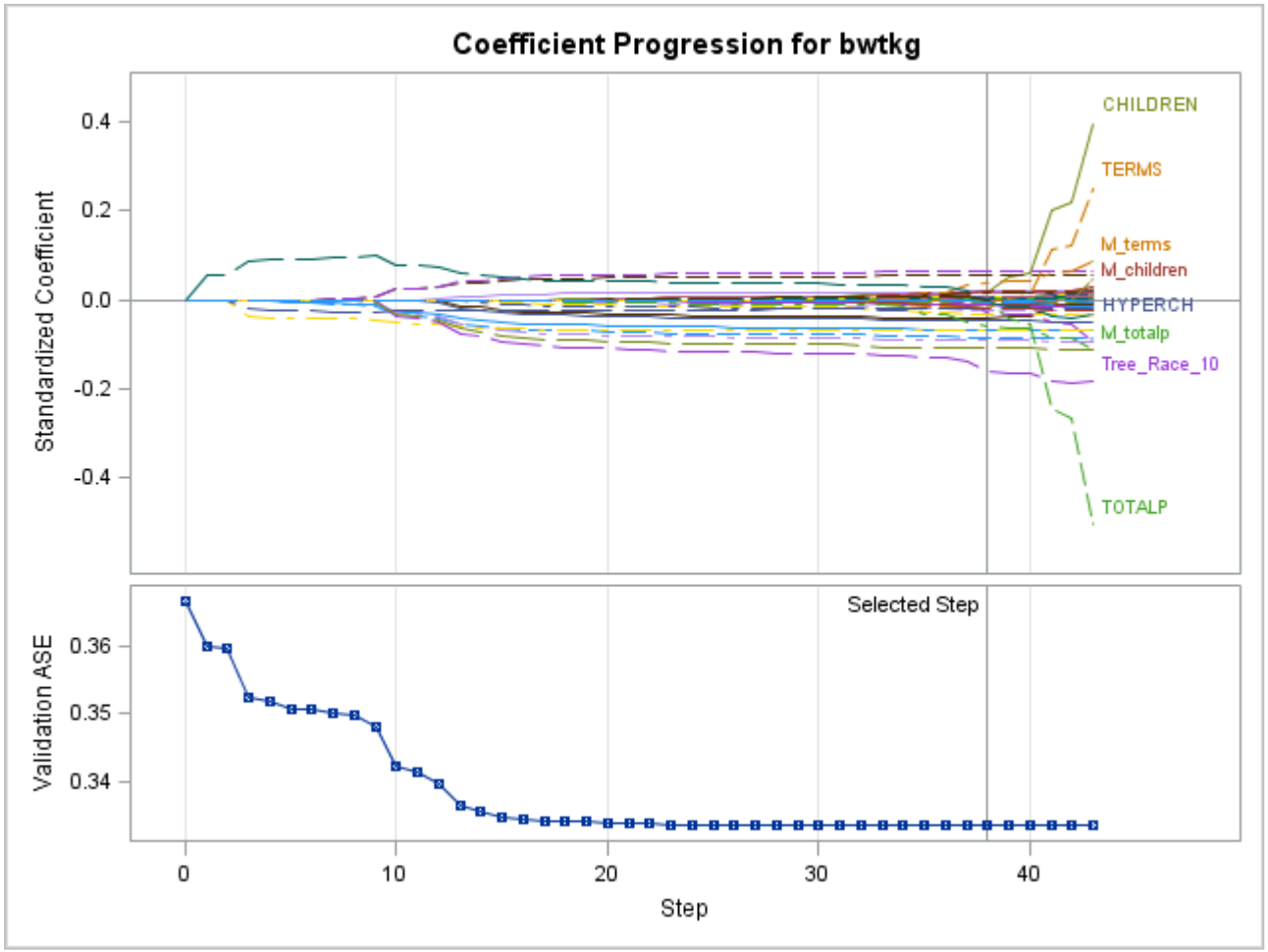
Backward Model Tuning using Validation ASE



Backward Model: Select and Choose using SBC



LAR Model Tuning using Validation ASE



Assess Final Model

```
title1 'BACKWARD: select SL choose validate';
ods html;
ods graphics on;
proc glm data=train plots(maxpoints=500000)=diagnostics;
    class tree_race;
    model bwtkg= HYPERCH CHILDREN CIGNUM LOUTCOME
        MEDUC PINFANT PRETERM TOTALP smoker Tree_Race/solution;
run;
quit;
```

- GLMSELECT does not provide model diagnostics.
- The model selected by GLMSELECT can be refit in PROC GLM.
- PLOTS=DIAGNOSTICS requests diagnostic plots. With larger datasets the user may have to increase the number of allowable plotting points using the MAXPOINTS= option.

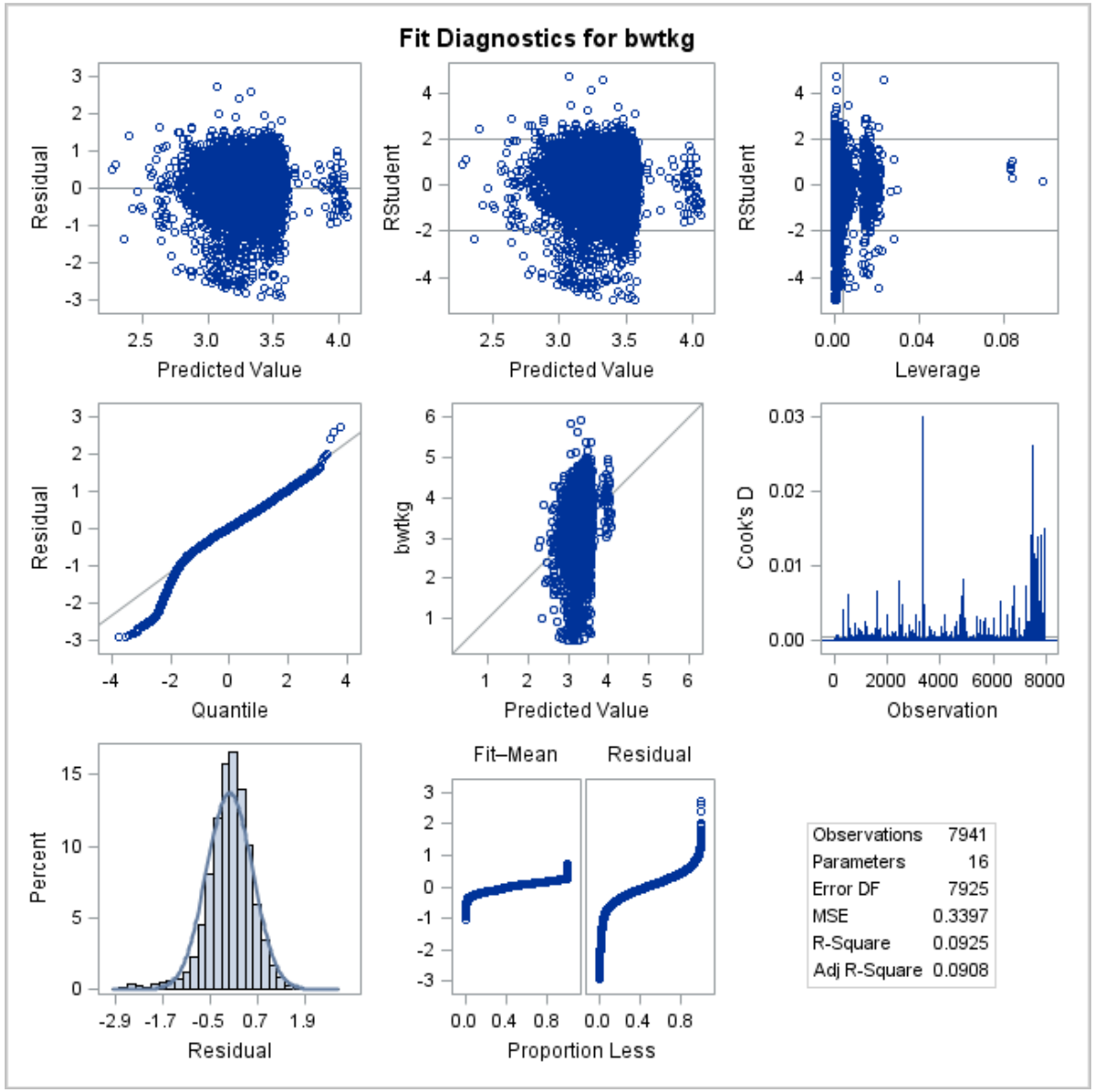
PROC GLM Model Estimates

- Tuning a model on Validation data error, especially when using a backward regression does not guarantee that all terms in the final model will be significant at the 5% level.

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	3.393074829	B	0.09321955	36.40	<.0001
HYPERCH	-0.316437355		0.07078175	-4.47	<.0001
CHILDREN	0.176907601		0.03535468	5.00	<.0001
CIGNUM	-0.009032741		0.00241774	-3.74	0.0002
DIABETES	0.063329386		0.04046288	1.57	0.1176
FEDUC	0.004833165		0.00356407	1.36	0.1751
LOUTCOME	0.043116199		0.02452329	1.76	0.0788
MEDUC	0.012849662		0.00346726	3.71	0.0002
PINFANT	0.470098794		0.07578353	6.20	<.0001
PRETERM	-0.538295956		0.06830569	-7.88	<.0001
TERMS	0.149955661		0.03661426	4.10	<.0001
TOTALP	-0.166673935		0.03451441	-4.83	<.0001
YrsLastLiveBirth	-0.005594258		0.00285376	-1.96	0.0500
drinker	-0.163389112		0.07007677	-2.33	0.0197
M_bdead	-0.577327101		0.36256303	-1.59	0.1113
M_children	0.883328150		0.42624567	2.07	0.0383
M_terms	1.225645880		0.40683507	3.01	0.0026
M_totalp	-1.396395899		0.41368468	-3.38	0.0007
M_YrsLastLiveBirth	-0.116547142		0.01980040	-5.89	<.0001
smoker	-0.124818184		0.03223567	-3.87	0.0001
MARITAL	-0.027463817		0.01855666	-1.48	0.1389
Tree_Race 6	-0.087179833	B	0.11168668	-0.78	0.4351
Tree_Race 7	0.049543737	B	0.07514436	0.66	0.5097
Tree_Race 8	0.013329560	B	0.07930082	0.17	0.8665
Tree_Race 9	0.053123518	B	0.08561883	0.62	0.5350
Tree_Race 10	-0.209024226	B	0.07644620	-2.73	0.0063
Tree_Race 11	-0.222655587	B	0.07842122	-2.84	0.0045
Tree_Race 99	0.000000000	B	.	.	.

PROC GLM Statistical Graphics Diagnostics

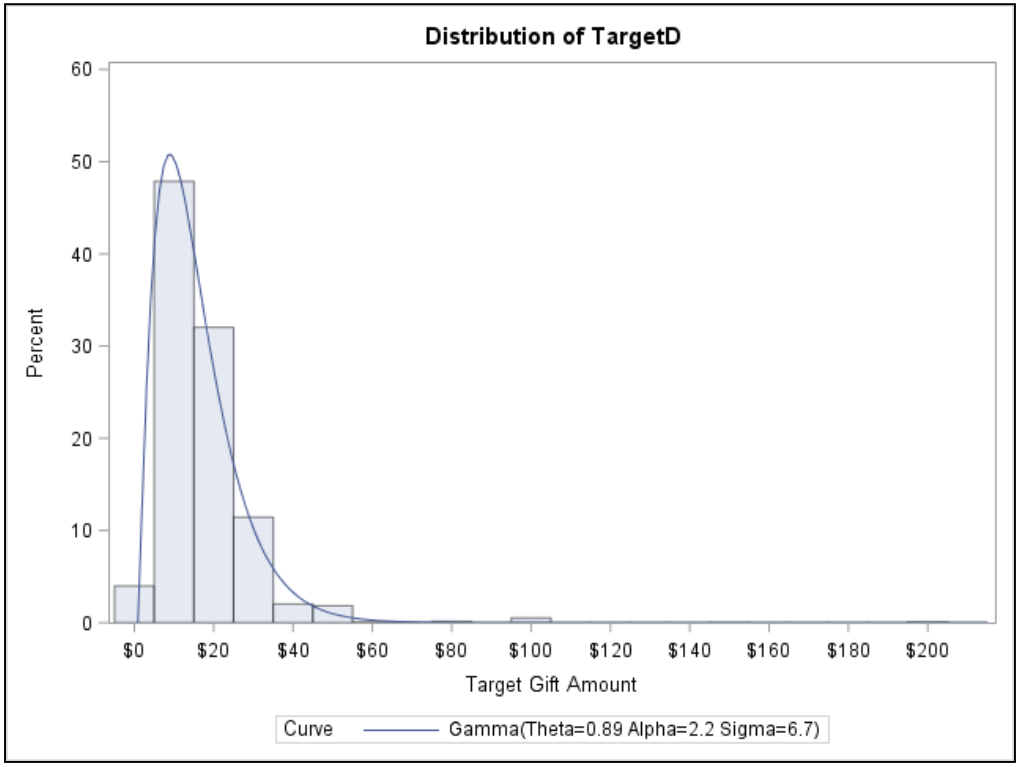
- ODS GRAPHICS ON and PLOTS=DIANGOSTICS.



Generalize Linear Predictive Modeling

The Gamma Distribution for Right-Skewed Targets: Charity Donation Data

```
title "Gamma Distribution";
proc univariate data=dev;
var targetd;
histogram /gamma(alpha=est sigma=est theta=est color=blue w=2)
    midpoints=0 to 210 by 10;
run;
```



- Donation amount, as is often the case for monetary or usage outcomes, is right skewed.
- An alternative to the Lognormal is the Gamma Distribution from the Exponential Family.

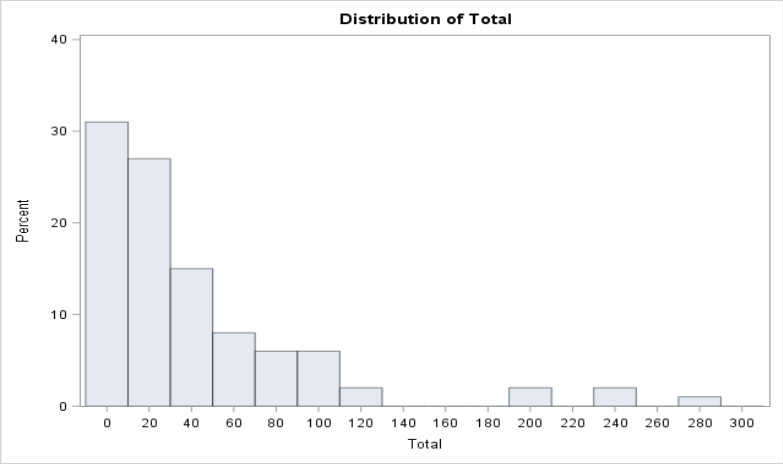
New Modeling Routines in SAS/STAT: Generalized Linear Models and Mixture Distributions using HPGENSELECT

- Fits Generalized Linear Models by specifying a distribution and link function to enable modeling of count data, rates, and non-normal continuous outcomes.

- Supports model selection routines
  - Backward, Forward, Stepwise selection using significance level
  - LASSO selection
  - Choice of final model using significance level, AIC, SBC

- Supports mixture distributions
  - Zero Inflated Poisson
  - Zero Inflated Negative Binomial
  - Tweedie

Mixture Models with HPGENSELECT



- The Tweedie distribution has been used extensively in insurance data modeling as it corresponds to the underlying loss generating process (e.g. total cost of claims). This mixed type distribution results from the mixing of two underlying components - the Poisson and Gamma distributions.

- In the Zero-Inflated Poisson Model, the population is considered to consist of two types of individuals. The first type gives Poisson distributed counts, which might contain zeros. The second type always gives a zero count. The ZIP model fits, simultaneously, two separate regression models. One is a logistic model that models the probability of being eligible for a non-zero count. The other models the size of that count.

Fit a Gamma Regression to the Charity Donation Data

```
proc hpgenselect data=dev;
     class &catvars;
     partition fraction(validate=.33 seed=12345);
     selection method=backward(slstay=.0001 choose=sbc);
     model targetd = &demvars &loggiftvars &cntvars
              &timevars &promvars &catvars/dist=gamma link=log;
run;
```

- A single development dataset is supplied to the procedure. The PARTITION statement requests a 67/33% split of this set.
- A backward regression is run, eliminating terms based on statistical significance. The final model in the backward sequence is chosen using the Schwartz Bayesian Criterion.
- A Gamma distribution is fit using a log link (though inverse is the canonical link).

DISTRIBUTION=	Distribution Function
BINARY	Binary
BINOMIAL	Binary or binomial
GAMMA	Gamma
INVERSEGAUSSIAN | IG	Inverse Gaussian
MULTINOMIAL | MULT	Multinomial
NEGATIVEBINOMIAL | NB	Negative binomial
NORMAL | GAUSSIAN	Normal
POISSON	Poisson
TWEEDIE<(*Tweedie-options*)>	Tweedie
ZINB	Zero-inflated negative binomial
ZIP	Zero-inflated Poisson

LINK=	Link Function	$g(\mu) = \eta =$
CLOGLOG | CLL	Complementary log-log	$\log(-\log(1-\mu))$
GLOGIT | GENLOGIT	Generalized logit	
IDENTITY | ID	Identity	μ
INV | RECIP	Reciprocal	$\frac{1}{\mu}$
INV2	Reciprocal square	$\frac{1}{\mu^2}$
LOG	Logarithm	$\log(\mu)$
LOGIT	Logit	$\log(\mu/(1-\mu))$
LOGLOG	Log-log	$-\log(-\log(\mu))$
PROBIT	Probit	$\Phi^{-1}(\mu)$

HPGENSELECT Output

Selection Summary

Step	Effect Removed	Number Effects In	SBC	p Value
0		23	19675.7499	.
1	DemHomeOwner	22	19667.6530	0.9776
2	DemPctVeterans	21	19659.6797	0.7250
3	GiftCntCardAll	20	19651.9043	0.5687
4	DemGender	19	19637.1617	0.4824
5	DemMedHomeValue	18	19629.8058	0.3885
6	PromCntCardAll	17	19622.7255	0.3099
7	PromCntAll	16	19615.2464	0.4354
8	GiftTimeLast	15	19607.9796	0.3586
9	DemAge	14	19600.8299	0.3332
10	log_GiftAvg36	13	19594.1801	0.2352
11	GiftCntAll	12	19588.3280	0.1373
12	PromCnt36	11	19586.2952	0.0139
13	PromCnt12	10	19579.5782	0.2369
14	GiftCntCard36	9	19575.7870	0.0401
15	GiftTimeFirst	8	19573.2257*	0.0189
16	GiftCnt36	7	19575.9507	0.0010

* Optimal Value of Criterion

- The model at step 15 minimizes SBC and is selected.

Fit Statistics

	Training	Validation
-2 Log Likelihood	19468	9748.57
AIC (smaller is better)	19494	9774.57
AICC (smaller is better)	19494	9774.80
BIC (smaller is better)	19573	9844.12
Pearson Chi-Square	492.72	437.83
Pearson Chi-Square/DF	0.1504	0.2836
Average Square Error	64.5186	127.86

- Fit Statistics are given for Training and Validation data.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.008108	0.052620	0.0237	0.8775
log_GiftAvgAll	1	0.464903	0.029182	253.8090	<.0001
log_GiftAvgCard36	1	0.281304	0.026679	111.1762	<.0001
log_GiftAvgLast	1	0.302822	0.018331	272.8930	<.0001
GiftCnt36	1	-0.015010	0.004545	10.9068	0.0010
PromCntCard12	1	0.034080	0.006354	28.7715	<.0001
PromCntCard36	1	-0.014484	0.002462	34.6165	<.0001
StatusCat96NK A	1	0.034258	0.016884	4.1172	0.0424
StatusCat96NK E	1	0.106647	0.045467	5.5017	0.0190
StatusCat96NK F	1	-0.154471	0.036044	18.3666	<.0001
StatusCat96NK L	1	0.028124	0.111557	0.0636	0.8010
StatusCat96NK N	1	-0.072344	0.032554	4.9386	0.0263
StatusCat96NK S	0	0	.	.	.
Dispersion	1	7.525182	0.181657	.	.

- Estimates and significance tests are provided.

Thank You

Lorne.Rothman@sas.com

THE POWER TO KNOW®