Predictive Modeling using SAS

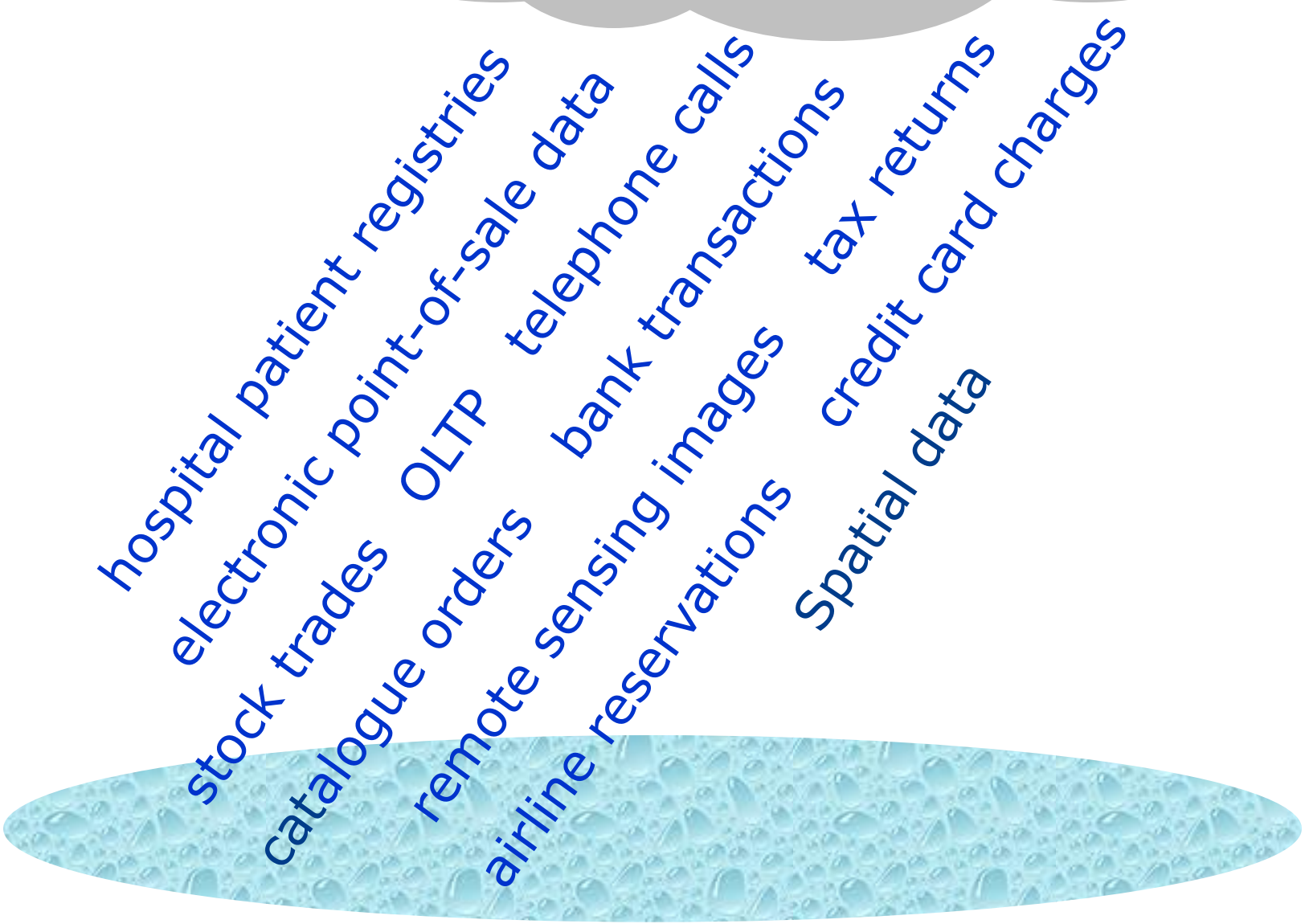
Purpose of Predictive Modeling

- ✓ 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

Challenges: Opportunistic Data

	Experimental	Opportunistic
Purpose	Research	Operational
Value	Scientific	Commercial
Generation	Actively controlled	Passively observed
Size	Small	Massive
Hygiene	Clean	Dirty
State	Static	Dynamic

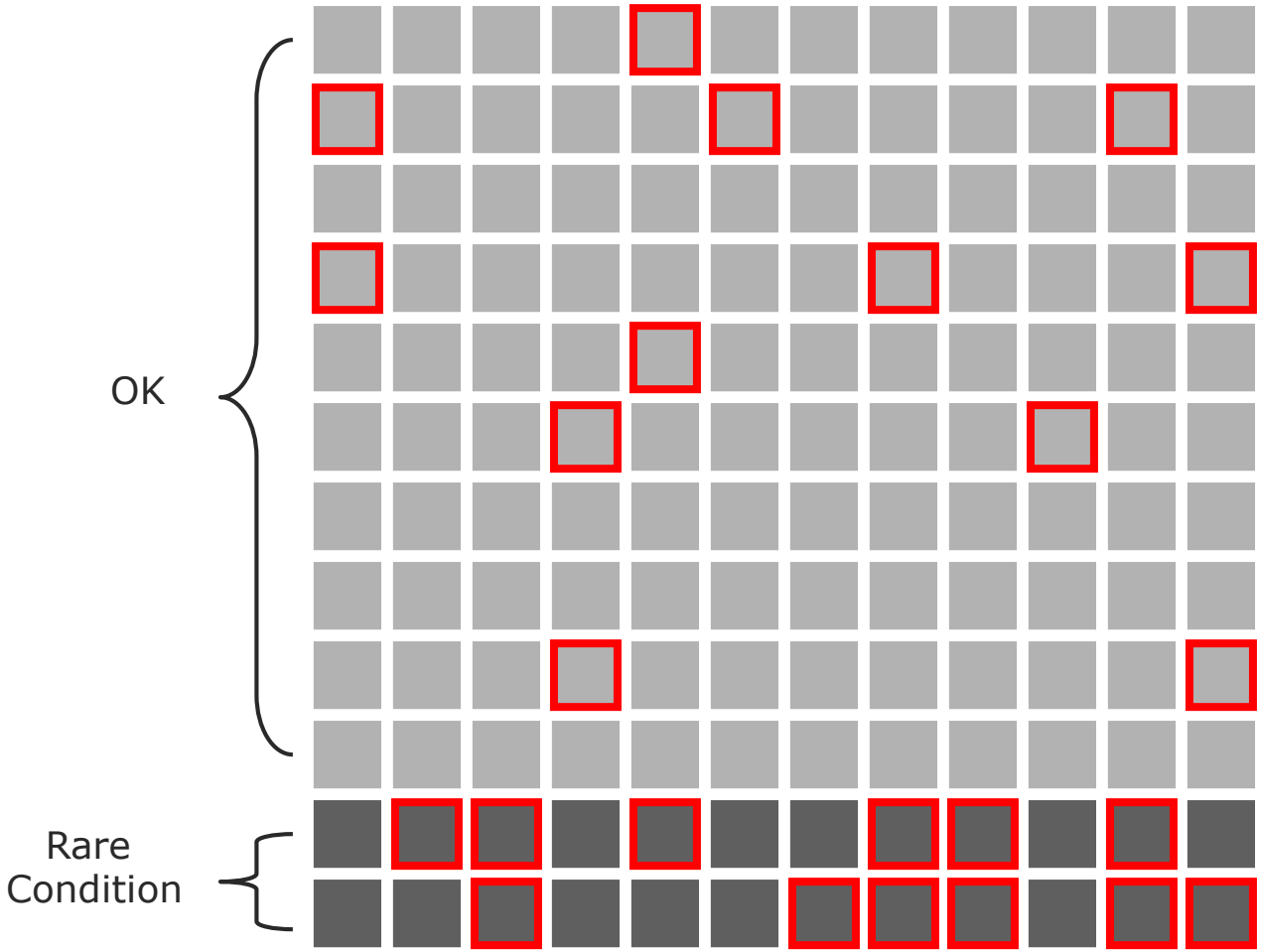
Challenges: Data Deluge



Challenges: Errors, Outliers, and Missings

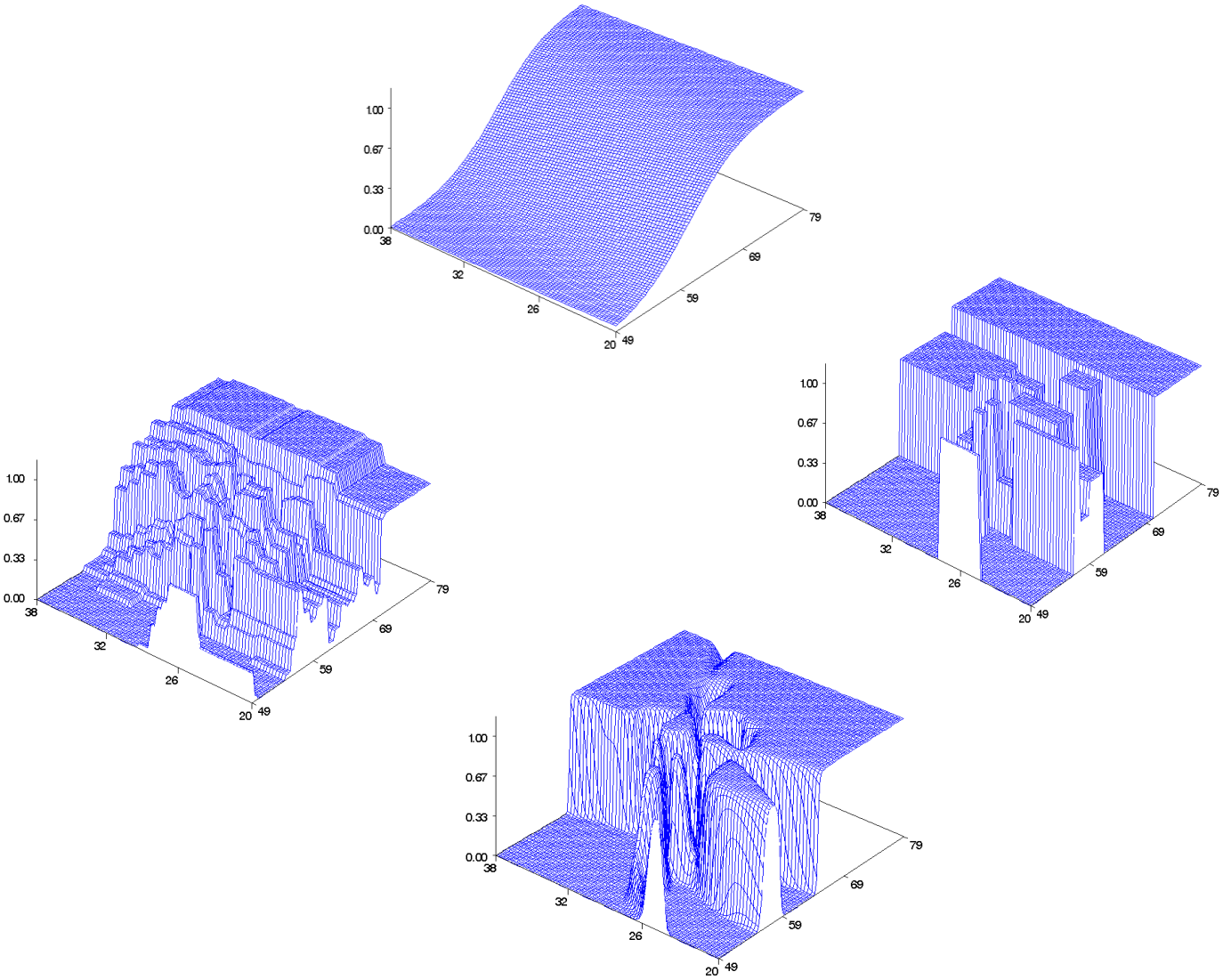
cking	#cking	ADB	NSF	dirdep	SVG	bal
Y	1	468.11	1	1876	Y	1208
Y	1	68.75	0	0	Y	0
Y	1	212.04	0	6		0
	.	.	0	0	Y	4301
y	2	585.05	0	7218	Y	234
Y	1	47.69	2	1256		238
Y	1	4687.7	0	0		0
	.	.	1	0	Y	1208
Y	.	.	.	1598		0
	1	0.00	0	0		0
Y	3	89981.12	0	0	Y	45662
Y	2	585.05	0	7218	Y	234

Challenges: Rare Events



Methodology: Empirical Validation

Methodology: Diversity of Algorithms



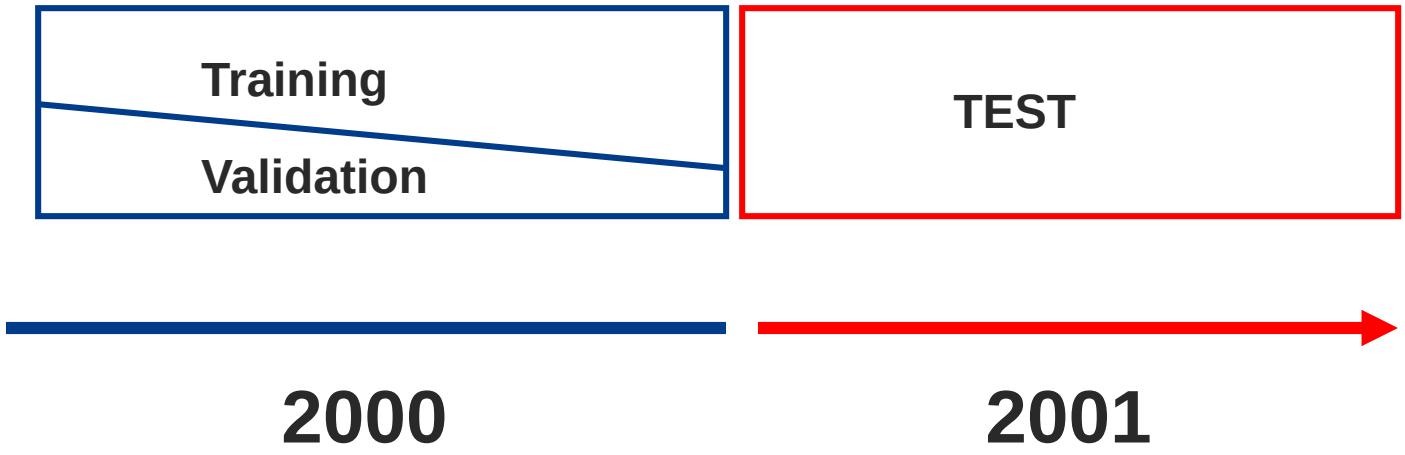
Jargon…

- 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.
- Data Scientist = What someone who used to be a data miner and before that a statistician calls themselves when looking for a job.

Binary Target Example: Predicting Low Birth Weight

- North Carolina Birth Records from North Carolina Center for Health Statistics
- 7.2% low birth weight births (< 2500 grams) excluding multiple births
- An oversampled (50% LBWT) development set of 17,063 births from 2000 and test set of 16,656 births from 2001
- Data contains Information on parents ethnicity, age, education level and marital status
- Data contains information on mothers health condition and reproductive history.

Predicting the Future with Data Splitting



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

Scenario: an early warning system for LBWT

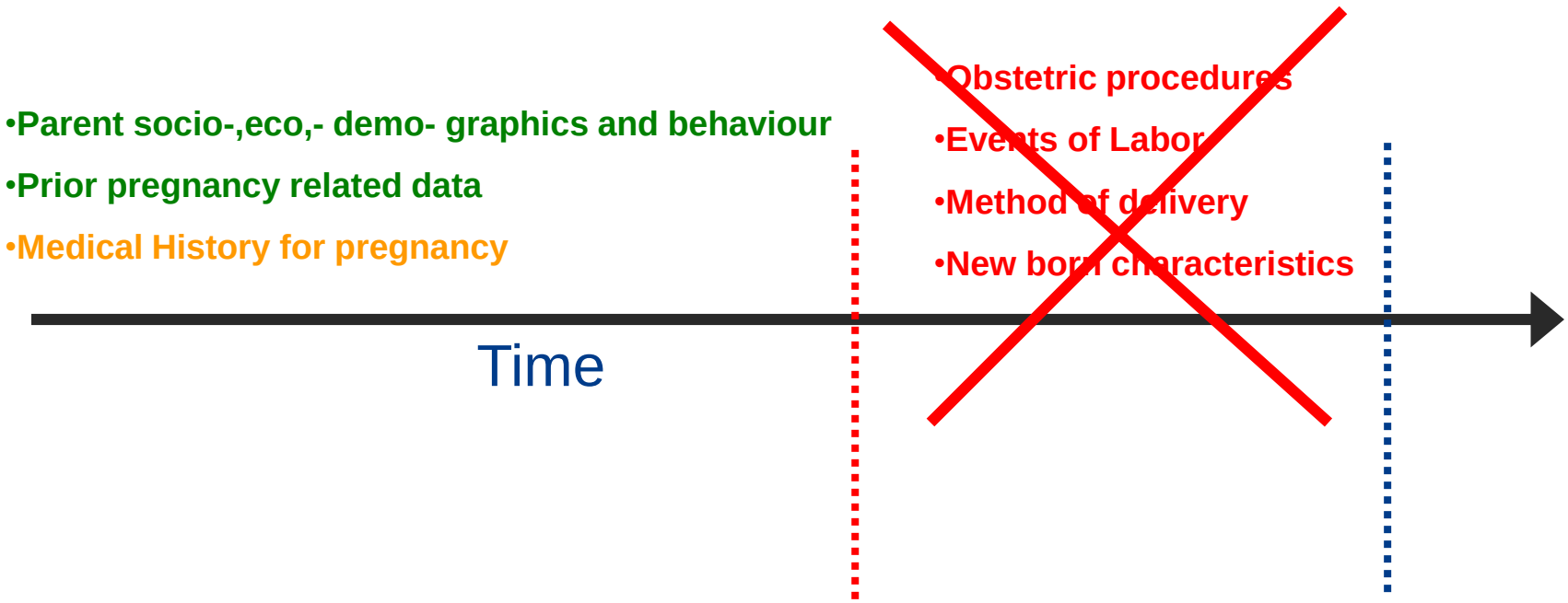
PREDICTORS

- **Parent socio-,eco-, demo- graphics, health and behaviour**
  - **Age, edu, race, medical conditions, smoking etc.**
- **Prior pregnancy related data**
  - **# pregnancies, last outcome, prior pregnancies etc.**

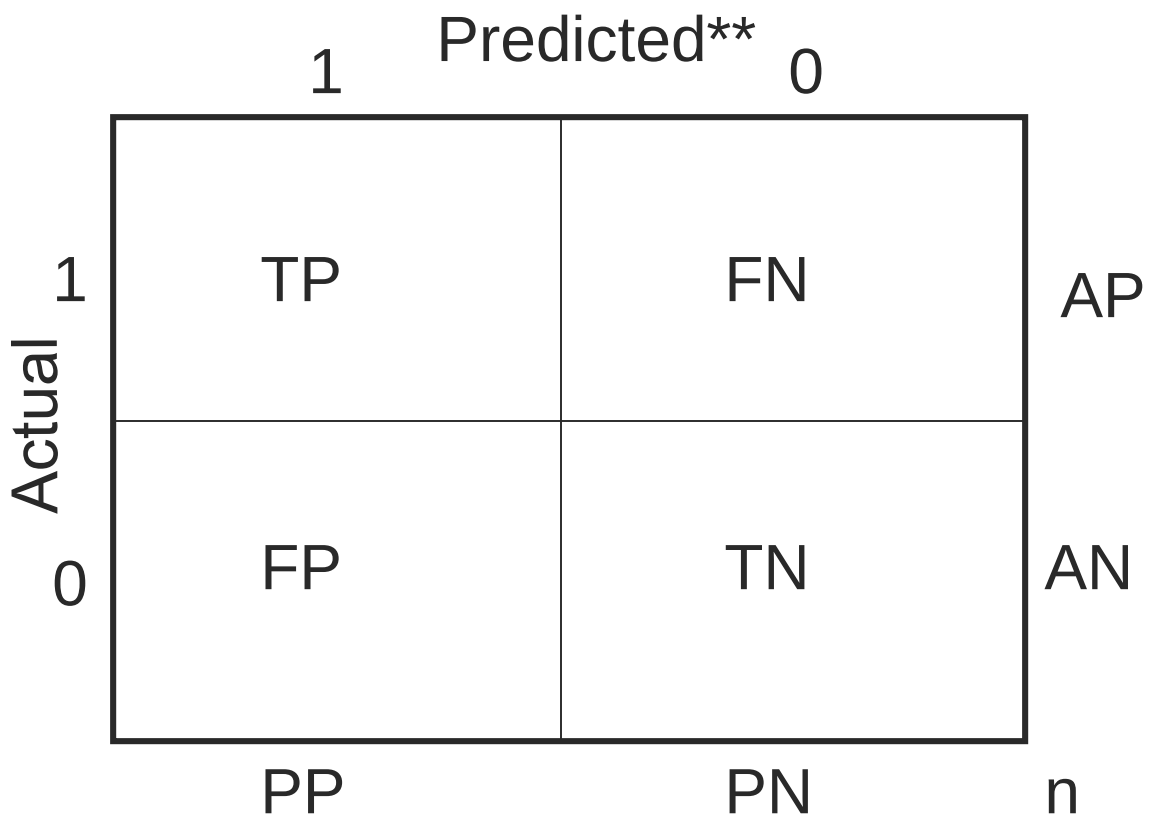
- **Medical History for pregnancy**
  - **Hypertension, cardiac disease, etc.**
- **Obstetric procedures**
  - **Amniocentesis, ultrasound, etc.**

- **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…..



Model Assessments for Binary Targets



Accuracy = (TP+TN)/n

Sensitivity = TP/AP

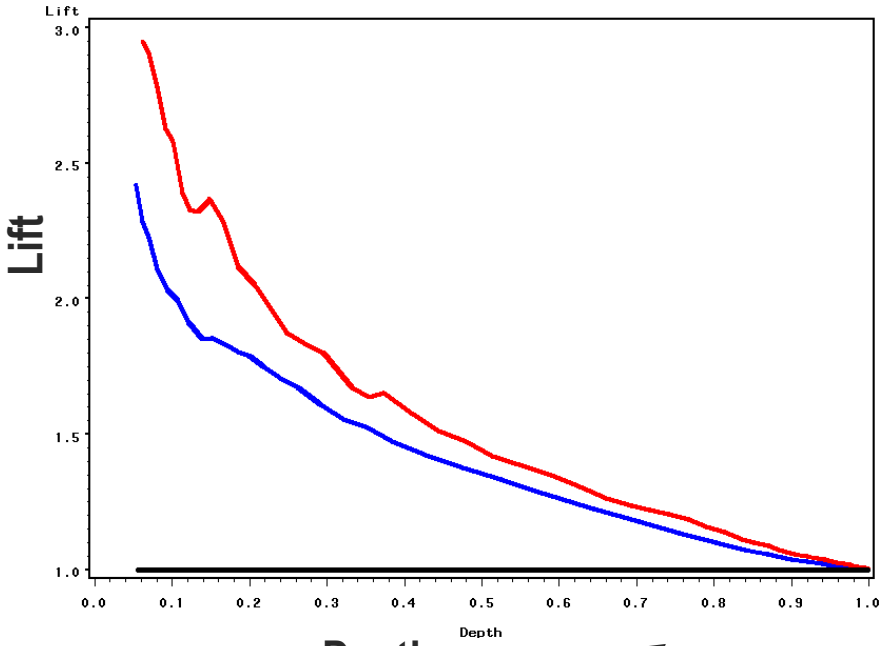
Specificity = TN/AN

Lift = (TP/PP)/π_1

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

Assessment Charts for Binary Targets

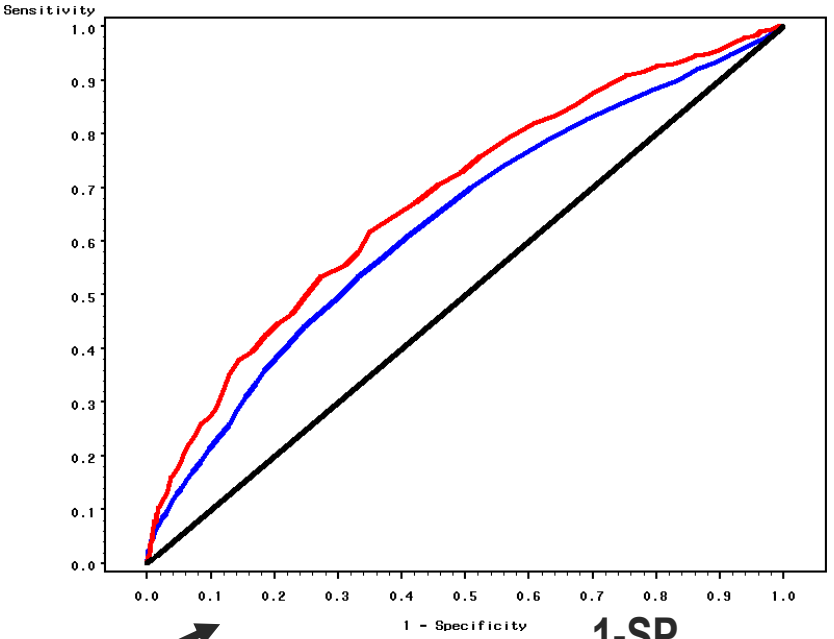
Lift Charts



Lift

Depth

ROC Charts



SE

1-SP

Explore measures across a range of cutoffs

TP	FN
FP	TN

TP	FN
FP	TN

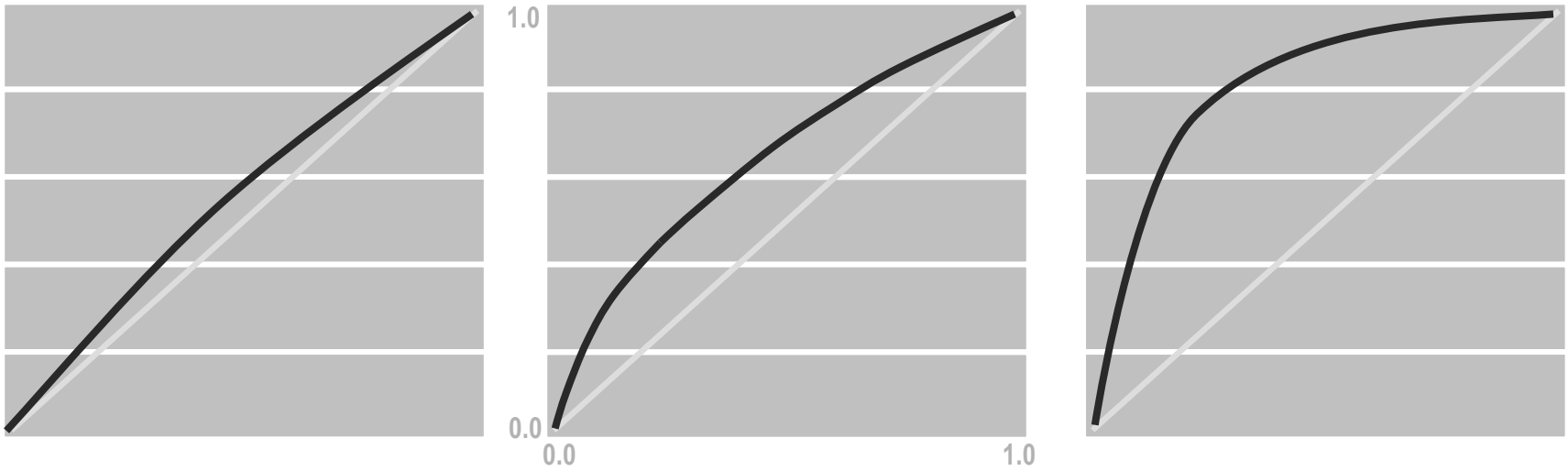
TP	FN
FP	TN

TP	FN
FP	TN

TP	FN
FP	TN

TP	FN
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).

Key Features of SAS STAT Code: Data Partition

```
proc surveyselect
                    data=pm.dev00
                    samprate=.6667
                    out=dev00
                    seed=44444
                    outall;
                    strata lbwt;
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.
- Stratification on the target, LBWT ensures equal representation of low birth weight cases in training and validation sets.

Key Features of SAS STAT Code: Imputation

```
proc stdize data=train reponly method=median
        out=train outstat=med;
   var _numeric_;
run;

proc stdize data=valid out=valid
            reponly method=in(med);
var _numeric_;
run;

proc stdize data=pm.test01 out=test
            reponly method=in(med);
var _numeric_;
run;
```

- STDIZE will do missing value replacement (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.

Key Features of SAS STAT Code

```
proc logistic data=train noprint;
   class &classvars;
   model lbwt(event='1')=&all;
   score data=valid out=sco_validate(rename=(p_1=p_all)) priorevent=.072;
run;

proc logistic data=train noprint;
 class &classvars;
   model lbwt(event='1')=&allint;
   score data=sco_validate out=sco_validate(rename=(p_1=p_AllInt))priorevent=.072;
run;

proc logistic data=train noprint;
 class &classvars;
   model lbwt(event='1')=&early;
   score data=sco_validate out=sco_validate(rename=(p_1=p_early))priorevent=.072;
run;
```

- After selecting three final models using stepwise methods, these three models are fit in LOGISTIC.
- The SCORE statement allows for scoring of new data and adjusts oversampled data back to the population prior (PRIOREVENT=0.072).
- The same dataset is re-scored (Sco_validate) so that predictions for all three models are in the same set for comparisons.
- The process is repeated using the Test set.

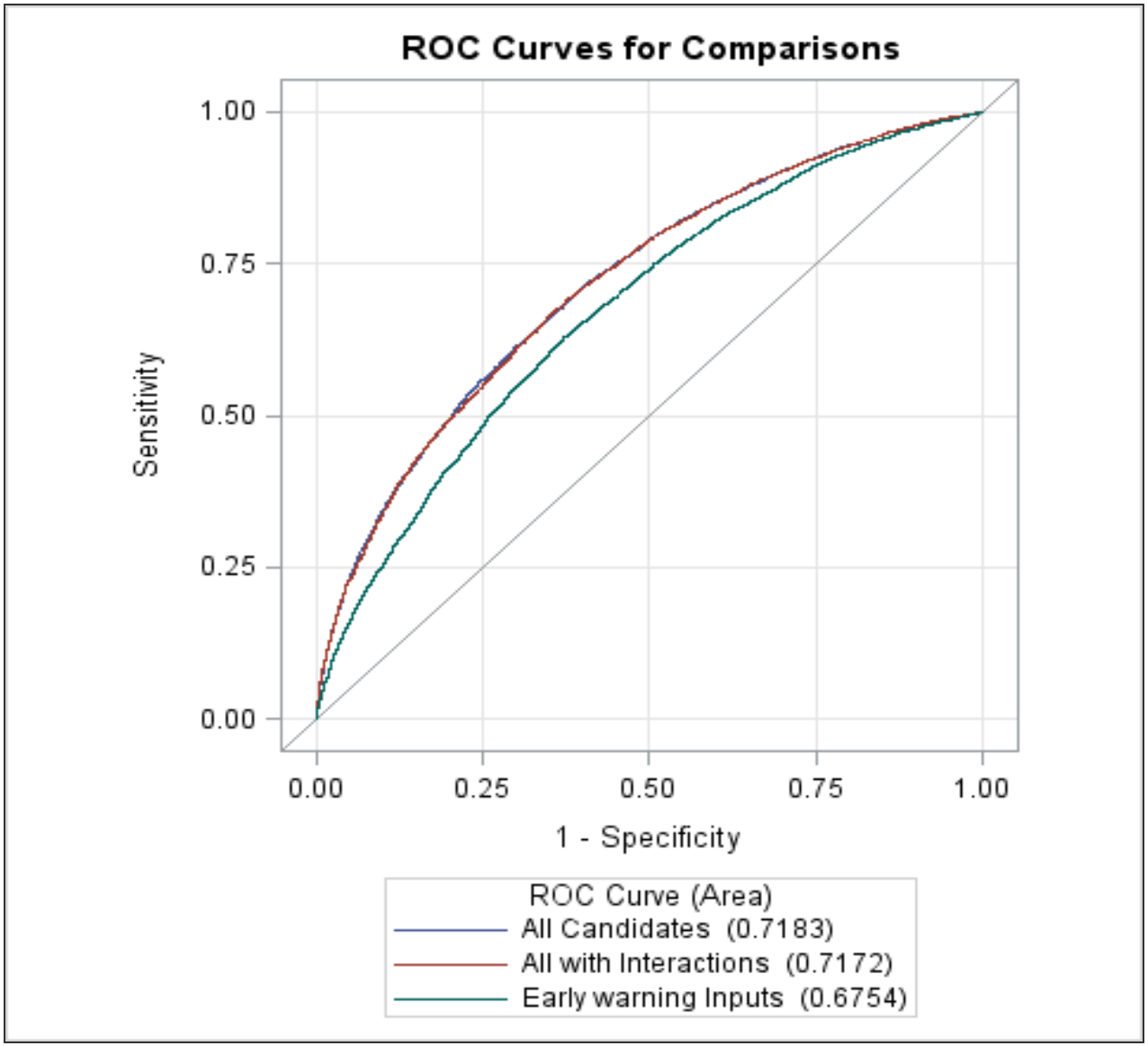
Key Features of SAS STAT Code

```
ods graphics on;
proc logistic data=sco_validate;
   model lbwt(event='1')=p_all p_allint p_early / nofit;
   roc "All Candidates" p_all;
   roc "All with Interactions" p_allInt;
   roc "Early warning Inputs" p_early;
   roccontrast "Comparing the Three Models: Validation Data "/estimate=allpairs;

run;
```

- The dataset with all three predictions (Sco_validate) is supplied to PROC LOGISTIC.
- The ROCCONTRAST statements provides statistical significance 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 process is repeated with the Test set.

Comparing ROC curves



Comparing ROC curves

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
All Candidates	0.7183	0.00391	0.7106	0.7259	0.4365	0.4367	0.2183
All with Interactions	0.7172	0.00392	0.7095	0.7248	0.4343	0.4345	0.2172
Early warning Inputs	0.6754	0.00412	0.6673	0.6834	0.3507	0.3508	0.1754

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Comparing the Three Models: Test Data	2	304.3867	<.0001

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
All Candidates - All with Interactions	0.00110	0.000576	-0.00003	0.00223	3.6207	0.0571
All Candidates - Early warning Inputs	0.0429	0.00248	0.0380	0.0478	299.4384	<.0001
All with Interactions - Early warning Inputs	0.0418	0.00256	0.0368	0.0468	267.5383	<.0001

DEMONSTRATION

THE POWER TO KNOW.®

Interval Target Example: Predicting Donation Amounts

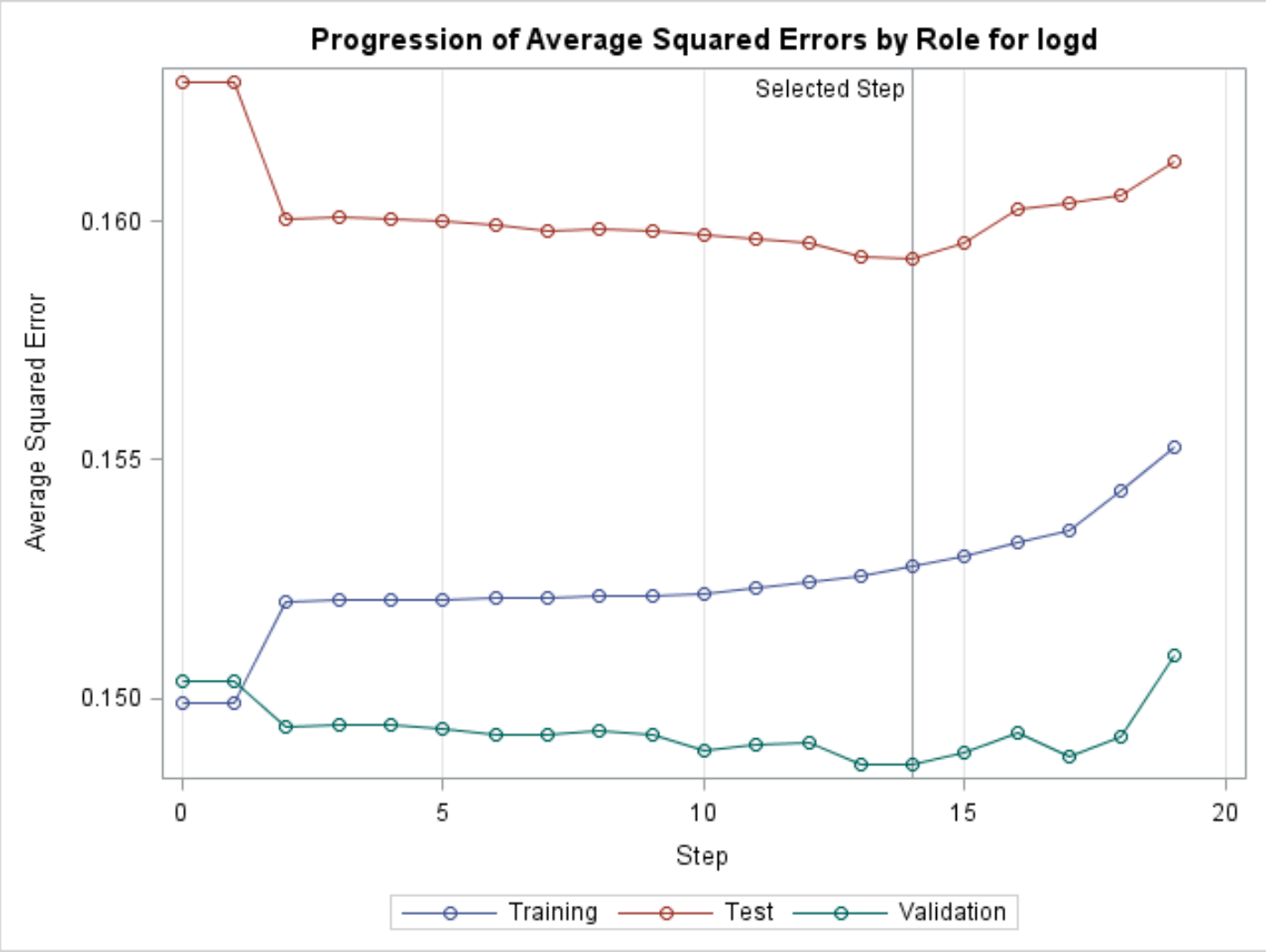
- A veterans' organization seeks continued contributions from lapsing donors. Use lapsing-donor donation amounts from an earlier campaign to predict future donations.
- Inputs include information on previous donation behavior by donors and solicitations by the charity.
- For example…DEMVARS: socioeconomic/demographic information, GIFTVARS: donation amount attributes, CNTVARS: donation frequency information, PROMVARS: Solicitation frequencies.

Key Features of SAS STAT Code

```
ods graphics on;
proc glmselect data=train valdata=valid testdata=test
        plots(stepAxis=number)=ASEPlot;
    class &catvars;
    model &target =  &demvars &loggiftvars &cntvars &timevars &promvars &catvars
        /selection=backward(choose = validate select = sl slstay=.0000001);
run;
```

- GLMSELECT fits interval target models 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.
- 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.
- 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.

Model Tuning using Validation ASE



Final Model Fitting and Score Code in GLM

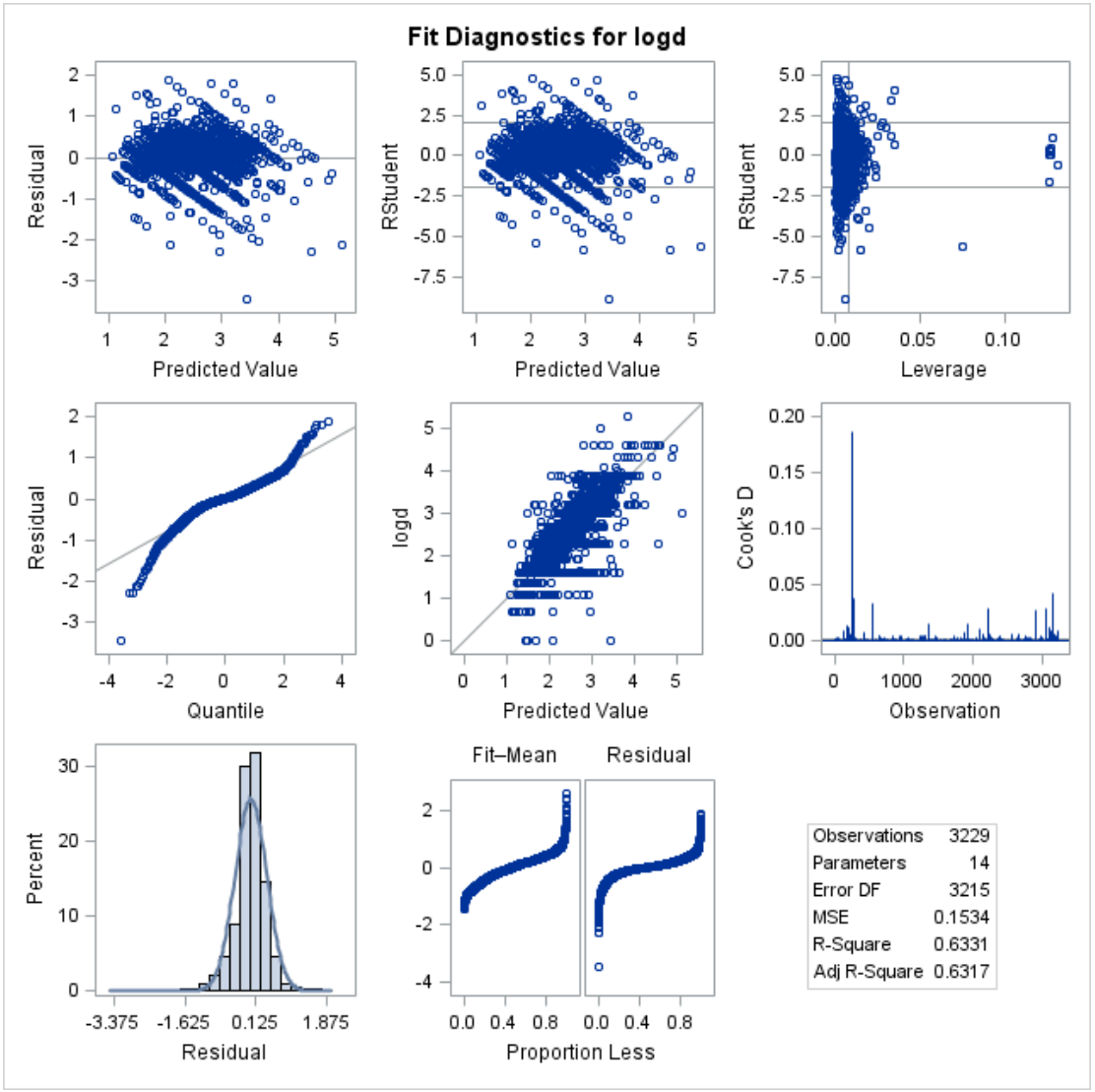
```
ods graphics on;
proc glm data=train plots=diagnostics;
class statuscat96nk;
model &target= log_GiftAvgAll log_GiftAvgCard36 log_GiftAvgLast GiftCnt36 PromCnt12
    PromCnt36 PromCntCard12 PromCntCard36 StatusCat96NK/solution;
    code file = 'C:\DATA\EDU\TALKS\UGwest2014\DonationMod.sas';
run;
quit;

data scored;
    set test;
    %include donationMod/source2;
run;
```

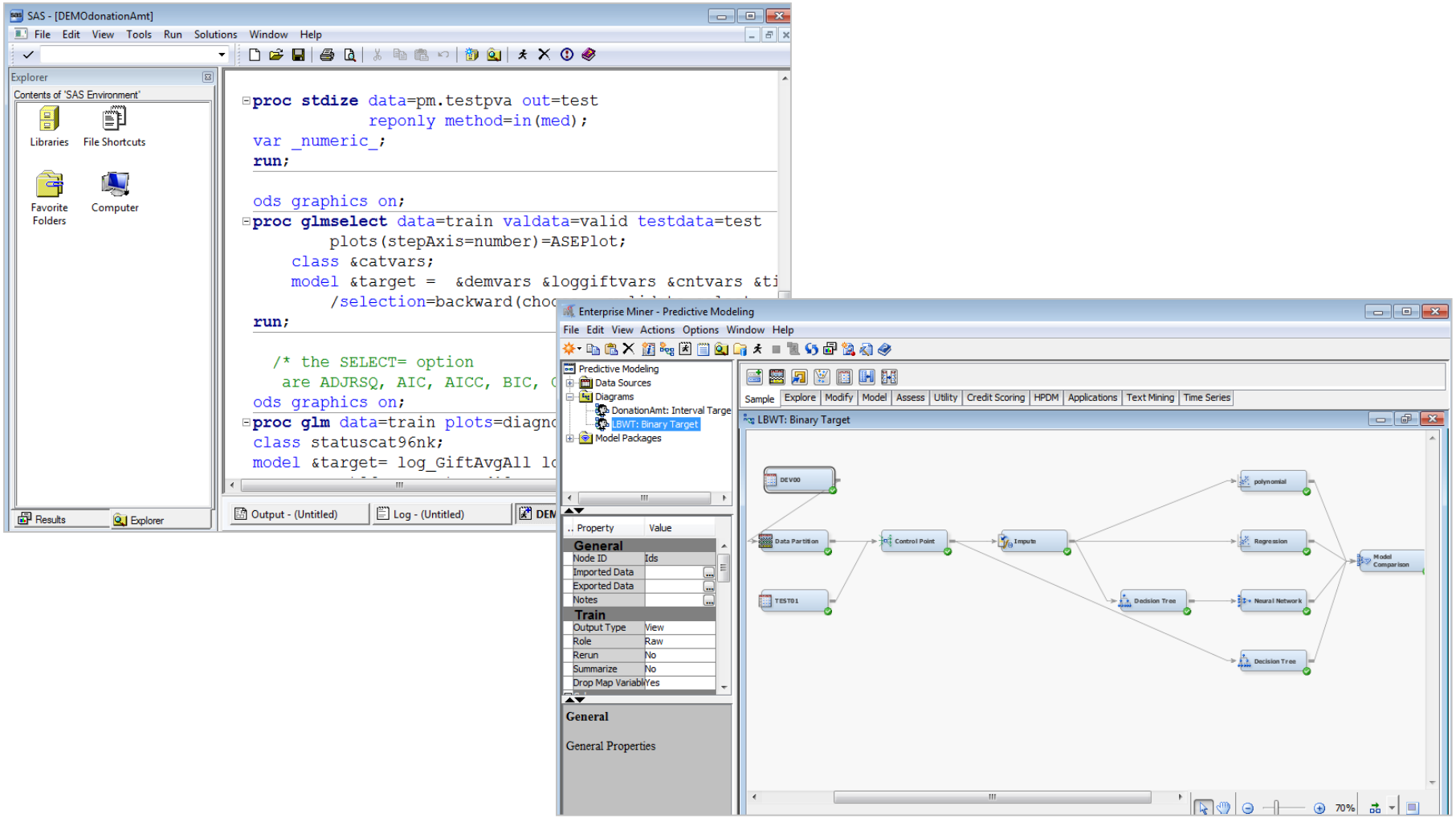
- GLMSELECT does not provide hypothesis test results and model diagnostics.
- The model selected by GLMSELECT can be refit in PROC GLM.
- PLOTS=DIAGNOSTICS requests diagnostic plots.
- The new CODE statement requests score code that can be applied to a new set with the %INCLUDE statement. SOURCE2 prints the scoring action to the log.
- The following procedures support a CODE statement as of V12.1: GENMOD, GLIMMIX, GLM, GLMSELECT, LOGISTIC, MIXED, PLM, and REG.

PROC GLM Statistical Graphics Diagnostics

- ODS GRAPHICS ON and PLOTS=DIANGOSTICS.



Predictive Modeling: Foundation SAS or Enterprise Miner



DEMONSTRATION

THE POWER TO KNOW.®

Thank You!

Lorne Rothman, PhD, P.Stat.
Principal Statistician
Lorne.Rothman@sas.com

THE POWER TO KNOW®