Simple and powerful: Predictive modeling using SAS/STAT

Wisconsin Illinois SAS Users Conference
June 25, 2014

Doug Thompson, PhD
Blue Cross Blue Shield of IL, MT, NM, OK & TX
Chicago, IL

Predictive Modeling

- Predictive modeling is essential for cost-efficient operations in many industries
  - Healthcare
  - Telecommunications
  - Insurance
  - Financial services
  - Many others
- Often, the goal of predictive modeling is to estimate the likelihood of specific events before they happen (e.g., avoidable hospital readmissions, product purchase, service retention, loan default), in order to be able to influence those events and achieve desirable outcomes

Examples

1. Predict likelihood of high-speed Internet churn
2. Predict likelihood of insurance policy lapse
3. Predict count of homeowners insurance claims caused by a hurricane projected to move through a specific area
4. Predict likelihood of having an elective surgery next year
5. Predict a person's total healthcare spend next year

Predictive Modeling vs. Other Modeling

- Although they share some procedures, predictive modeling and other modeling have different goals and techniques
  - Predictive modeling
  - Hypothesis testing
  - Explanatory analysis
- In predictive modeling, the point is typically to use the model output ("score") to take some action; what matters most is that predictions are accurate and generalizeable
- May intentionally violate important principles of other modeling (e.g., distributional assumptions, multicollinearity) and ignore central aspects of other modeling (e.g., p-values, confidence intervals, interpretation)

Tools

- Maybe because predictive modeling has become so ubiquitous and lucrative, a variety of predictive modeling tools have appeared in the marketplace
  - SAS Enterprise Miner (EM)
  - SPSS data mining software
  - Vendor proprietary tools
- Some of these tools are extremely expensive and the return on investment is questionable

A Cost-Effective Solution

- SAS/STAT offers a wide variety of predictive modeling capabilities at an affordable price
- May be supplemented with freeware for decision trees, neural networks and other algorithms not available in SAS/STAT (although I believe this is seldom necessary)
  - R and Weka are both free
- Some large companies that have SAS/EM available still use SAS/STAT as their "workhorse" for predictive modeling, supplemented by EM in specific situations

This Presentation

- Illustrate the process of building a predictive model using SAS/STAT
- The illustrated process may not result in the best possible model (can't be guaranteed with any predictive modeling technique, someone is always building a better mouse trap) -- but it will likely result in a pretty good model with solid business value
- Goal is to overview some basic steps and provide a sense of the process

Illustration

- Goal of predictive model
  - Predict likelihood of high healthcare spend (>$20K) next year
- Data
  - Medical Expenditure Panel Survey (MEPS)
  - Panel 14 (2009-2010)
  - Panel/longitudinal with 5 rounds of data collection for each participant
  - Complex survey design (e.g., weights, PSUs and strata) intentionally ignored in this illustration

Steps

1. Define scope, sample and variables
2. Split sample
3. Descriptive profiling
4. Bivariate screening
5. Variable clustering
6. Variable selection
7. Transformations
8. Interactions
9. Create final model
10. Evaluate on holdout

1. Define scope, sample and variables

- What will the model predict (**dependent variable**)?
  - Likelihood of having healthcare spend >$20K next year
- On what population will the model be used (**scope**)?
  - People age <65 with private insurance
- What variables can be used to make the prediction (**predictors**)?
  - Healthcare spend this year
  - Other health indicators
  - Demographics

totexpy2_gt20k	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6529	97.13	6529	97.13
1	193	2.87	6722	100

2. Split sample

- Build model on one sample, evaluate on a different sample

```
data training holdout;
set inscope;
if ranuni(34820)<0.67 then output training;
else output holdout;
run;
```

3. Descriptive profiling

- Describe each predictor by category of dependent variable
- `%macro profile_and_bivariate;`

	Next year Healthcare spend				
Predictor	**<= $20K**	**> $20K**	**Chi Sq**	**P-value**	**C-stat**
Family size	3.4	2.6	33.20	<.0001	0.64
Northeast	15.4%	11.2%	1.82	0.1769	0.52
Age (years)	33.6	46.8	68.99	<.0001	0.71
Female	51.2%	58.4%	2.55	0.1100	0.54
Married	46.1%	60.8%	10.51	0.0012	0.57
College degree	25.7%	34.4%	4.46	0.0347	0.54
Health	2.0	2.8	69.29	<.0001	0.67
Mental health	1.8	2.2	24.18	<.0001	0.62
Employed	61.9%	64.8%	0.44	0.5070	0.51

4. Bivariate screening

- May select p<0.15 for screening purposes (situation dependent)

```
%let indvars =
_FAMSY1 midwest _age2x
...
totexpy1;

%do i=1 %to 22;
%let curr_var = %scan(&indvars,&i);

proc logistic data=training descending;
model totexpy2_gt20k = &curr_var;
ods output FitStatistics=fits Association=assoc GlobalTests=test;
run;
```

5. Variable clustering

- Recommended: Apply a combination of statistical criteria and business sense

```
proc varclus data=training;
var
_FAMSY1 midwest _age2x female hispanic married
_hs_or_less college_degree
_RTHLTH2 _MNHLTH2 born_in_us have_usc usc_pcp
_totexpy1;
ods output RSquare=RSquare;
run;
```

6. Variable selection

- Try p-value, stepwise, backward

```
proc logistic data=training descending
 namelen=100;
model totexpy2_gt20k =
_FAMSY1 midwest _age2x female hispanic married
 hs_or_less college_degree _RTHLTH2 _MNHLTH2
 born_in_us have_usc usc_pcp totexpy1;
ods output parameterestimates=parms;
run;
```

7. Transformations

- Some alternatives to linear: Set of categories, log, inverse, square root (the possibilities are endless); splines are good too (take a little more coding)

```
array orig{*} TOTEXPY1 _AGE2X _RTHLTH2;
array nlogs{*} ln_TOTEXPY1 ln_AGE2X ln_RTHLTH2;
array sqrts{*} sqrt_TOTEXPY1 sqrt_AGE2X sqrt_RTHLTH2;
array invs{*} inv_TOTEXPY1 inv_AGE2X inv_RTHLTH2;

do i=1 to 3;
   nlogs{i}=log(orig{i}+0.01);
   sqrts{i}=sqrt(orig{i});
   invs{i}=1/(orig{i}+0.01);
end;
```

8. Interactions

- Can consider all possible n-way interactions or use decision trees (not available in SAS/STAT) – see my paper from MWSUG '12

```
proc logistic data=training descending namelen=100;
model totexpy2_gt20k =
MIDWEST|
ln_TOTEXPY1|
_AGE2X|
_RTHLTH2|
born_in_US @3 / selection=stepwise;
run;
```

9. Create final model

- Final model equation:

```
logit=-7.43772102+
MIDWEST*0.45793467+
ln_TOTEXPY1*0.06222569+
_AGE2X*0.06602300+
_RTHLTH2*-0.40267581+
ln_TOTEXPY1*_RTHLTH2*0.16587382+
_AGE2X*_RTHLTH2*-0.01515550;
score=1/(1+exp(-1*logit));
```

	Next year Healthcare spend	
Predictor	**<= $20K**	**> $20K**
Midwest	23.9%	33.6%
Age (years)	33.6	46.8
Health	2.0	2.8
Last yr spend	$2,850	$16,570

10. Evaluate on holdout

- I typically look at holdout c-statistic and decile analysis
- Easy to generate with macros
  - C-stat macro from Izrael et al SUGI presentation (275-28)

```
%macro wtc (ds = , outds = , weight = , depvar =) ;
```

  - I use my own macro for decile analysis

```
%macro weighted_decile(infile=,  score=, target=,
  seed=0, weight=);
```

10. Evaluate on holdout (cont'd)

Score Decile	Total Count	Target Count	Expected Count	Cum. % Target
1	222	32	28.5	47.1%
2	221	14	11.5	67.6%
3	221	4	7.6	73.5%
4	221	5	5.2	80.9%
5	221	5	3.4	88.2%
6	222	3	2.3	92.6%
7	221	3	1.5	97.1%
8	221	0	0.9	97.1%
9	221	0	0.5	97.1%
10	221	2	0.0	100.0%
Total	**2,212**	**68**		

Using the Model

- Typically hand off to IT to productionize the scoring process
- Refresh on a regular cycle
- Push scores out to business users
- Often score *rank* is the focus – the highest-scoring *n* individuals receive some intervention
- Recalibrate/rebuild periodically

Questions and Comments Are Welcome

Doug Thompson, PhD
Blue Cross Blue Shield of IL, MT, NM, OK & TX
300 E Randolph
Chicago, IL 60601
doug_thompson@bcbsil.com