



SouthEast SAS[®]
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008

Paper MPSF-073

Model Validity Checks In Data Mining: A Luxury or A Necessity?

Mike Speed & Simon Sheather
Department of Statistics
Texas A&M University

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.
® indicates USA registration. Other brand and product names are trademarks of their respective companies.

1



SouthEast SAS[®]
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008

Abstract

It has been stated that, because of vast amounts of data, the data miner does not have the luxury of verifying assumptions and hence checking for model validity. This paper uses two databases to investigate implications of not verifying assumptions and hence the validity of models. The first one is the Dominick's Finer Foods database (James M. Kilts Center, GSB, University of Chicago). This database consists of same store sales of 58 stores over 268 weeks. The second one is The Direct Mail Fundraising database. This paper demonstrates that ensuring that the fitted models for either continuous or binary outcomes are valid improves predictions, sometimes dramatically so. In these real examples considered in this talk, using valid models, greatly improved the results. Thus, the data miner may want to consider that the checking of assumptions and validity is not a luxury, but in fact a necessity.

KEYWORDS: data mining model validity covariance structure

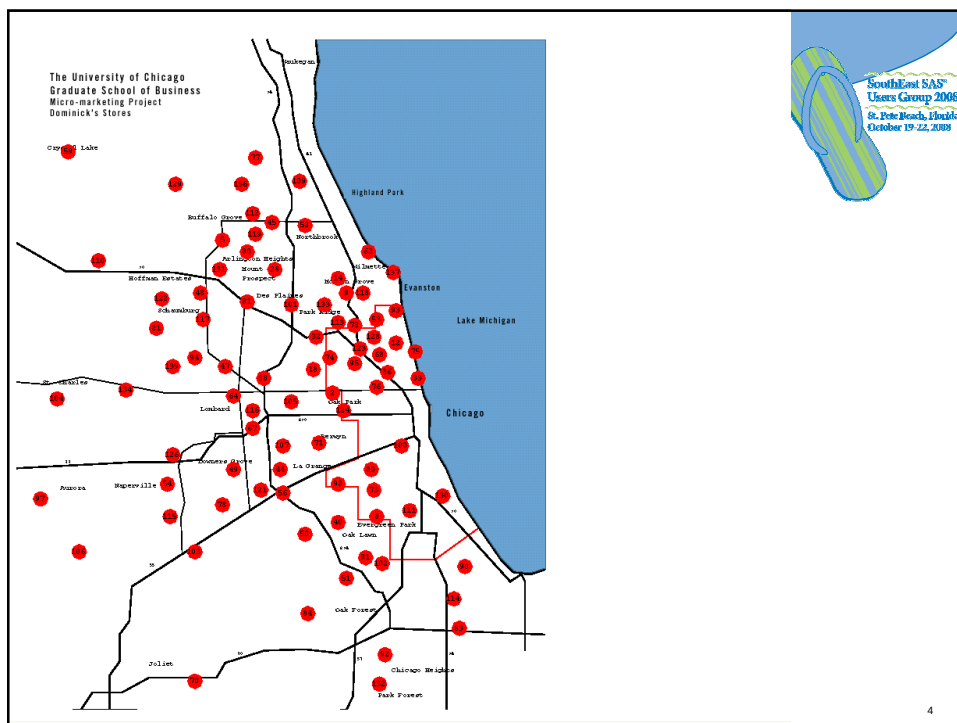
2

Database Number 1

**SouthEast SAS®
Users Group 2008**
St. Pete Beach, Florida
October 17-22, 2008

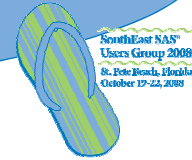
- **Dominick's: Customer Count Files**
- **James M. Kilts Center, GSB, University of Chicago**
- <http://research.chicagogsb.edu/marketing/databases/dominicks/index.aspx>

2



1

Dominick's Finer Foods



- **Dominick's: Customer Count Files**
- **James M. Kilts Center, GSB, University of Chicago**
- <http://research.chicagogsb.edu/marketing/databases/dominicks/index.aspx>

“The customer count file includes information about in-store traffic. The data is store specific and on a **daily** basis. The customer count data refers to the number of customers visiting the store and purchasing something. Also in the customer count file is a total dollar sales and total coupons redeemed figure, by DFF defined department. These figures are compiled daily from the register/scanner receipts.”

5

Data

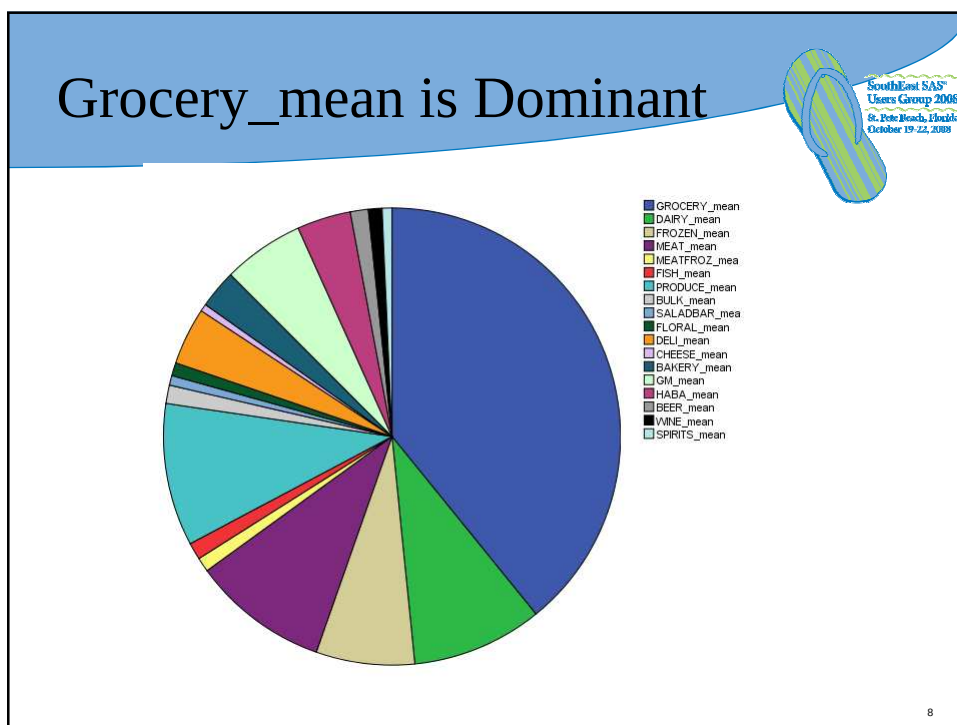


- Weekly averages were computed from the daily sales
- No coupon data was used
- Variables used: GROCERY_mean (Dependent)
DAIRY_mean; FROZEN_mean; MEAT_mean;
MEATFROZ_mean; FISH_mean ;PRODUCE_mean;
BULK_mean; SALADBAR_mean; FLORAL_mean;
DELI_mean; CHEESE_mean; BAKERY_mean; GM_mean;
HABA_mean; BEER_mean; WINE_mean; SPIRITS_mean;
CUSTCOUN_mean
- 58 stores
- 268 weeks
- 13998 observations

6

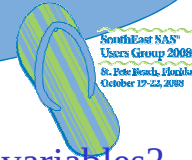
Variable	Description
GROCERY_mean	Grocery Sales in Dollars
Week	Week Number
Store	Store Code
BAKERY_mean	Bakery Sales in Dollars
BEER_mean	Beer Sales in Dollars
BULK_mean	Bulk Sales in Dollars
CHEESE_mean	Cheese Sales in Dollars
CONVFOOD_mean	Conventional Foods Sales in Dollars
CUSTCOUN_mean	Customer Count
DAIRY_mean	Dairy Sales in Dollars
DELI_mean	Deli Sales in Dollars
FISH_mean	Fish Sales in Dollars
FLORAL_mean	Floral Sales in Dollars
FROZEN_mean	Frozen Items Sales
GM_mean	General Merchandise Sales in Dollars
HABA_mean	Health and Beauty Aids Sales in Dollars
MEAT_mean	Meat Sales in Dollars
MEATFROZ_mean	Meat-Frozen Sales in Dollars
PRODUCE_mean	Produce Sales in Dollars
SALADBAR_mean	Salad Bar Sales in Dollars
SPIRITS_mean	Spirits Sales in Dollars
WINE_mean	Wine Sales in Dollars

7



8

Business Questions



SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008

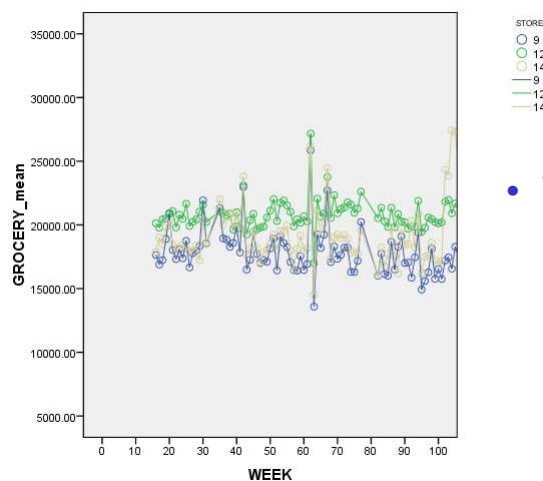
- How is Grocery_mean related to the other variables?
- Does increasing/decreasing the sales on one variable reduce or enhance the Grocery_mean sales?
- What role does customer count play on the Grocery_mean sales?
- Should we model the correlation among the same store sales?

9

Three Stores for Selected Weeks



SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008



- Use a 60/40 split
 - 60 is training set
 - 40 is validation set

10

Regression

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

Effect	Estimate	Pr > t
DAIRY_mean	1.80	<.0001
FROZEN_mean	1.35	<.0001
MEAT_mean	0.75	<.0001
MEATFROZ_mean	1.09	<.0001
FISH_mean	-1.60	<.0001
PRODUCE_mean	0.27	<.0001
BULK_mean	0.50	<.0001
SALADBAR_mean	0.50	<.0001
FLORAL_mean	0.04	0.5952
DELI_mean	0.98	<.0001
CHEESE_mean	-0.66	<.0001
BAKERY_mean	-0.42	<.0001
GM_mean	0.39	<.0001
HABA_mean	0.19	<.0001
BEER_mean	0.15	0.0577
WINE_mean	-0.78	<.0001
SPIRITS_mean	0.29	0.073
CUSTCOUN_mean	-0.68	<.0001

What seems
to be of interest here?

11

Regression

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

Fit Statistics

R-Square	0.9671	Adj R-Sq	0.9659
AIC	119040.2652	BIC	119062.4269
SBC	121052.5235	C(p)	286.0000

Mean Square Error 1382894

What seems
to be of interest here?

12

Regression

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

TARGET	Fit statistics	Statistics La...	Train	Validation	Test
GROCERY_m...	AIC_	Akaike's Infor...	119040.3	.	.
GROCERY_m...	ASE_	Average Squ...	1335804	1340604	.
GROCERY_m...	AVERR_	Average Erro...	1335804	1340604	.
GROCERY_m...	DFE_	Degrees of F...	8113	.	.
GROCERY_m...	DFM_	Model Degre...	286	.	.
GROCERY_m...	DFT_	Total Degree...	8399	.	.
GROCERY_m...	DIV_	Divisor for ASE	8399	5599	.
GROCERY_m...	ERR_	Error Function	1.122E10	7.506E9	.
GROCERY_m...	FPE_	Final Predicti...	1429984	.	.
GROCERY_m...	MAX_	Maximum Ab...	31345.01	11045.93	.
GROCERY_m...	MSE_	Mean Square...	1382894	1340604	.
GROCERY_m...	NOBS_	Sum of Frequ...	8399	5599	.
GROCERY_m...	NV_	Number of Es...	286	.	.
GROCERY_m...	RASE_	Root Averag...	1155.77	1157.845	.
GROCERY_m...	RFPE_	Root Final Pr...	1195.819	.	.
GROCERY_m...	RMSE_	Root Mean S...	1175.965	1157.845	.
GROCERY_m...	SBC_	Schwarz's B...	121052.5	.	.
GROCERY_m...	SSE_	Sum of Squa...	1.122E10	7.506E9	.
GROCERY_m...	SUMW_	Sum of Case ...	8399	5599	.

What seems
to be of interest here?

13

Run Regression As Proc Mixed -

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

- REPEATED WEEK / SUBJECT=STORE
TYPE=VC ;
- Same Results as with Regression

14

Run Mixed with TYPE = CS

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008

- REPEATED WEEK / SUBJECT=STORE
TYPE=CS ;

Compound
Symmetry

CS

$$\begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 \end{bmatrix}$$

- Here we have a 268 by 268 covariance matrix
- Different Results

15

Effect	Estimate		Pr > t	
	VC	CS	VC	CS
DAIRY_mean	1.80	<.0001	1.8030	<.0001
FROZEN_mean	1.35	<.0001	1.3086	<.0001
MEAT_mean	0.75	<.0001	0.4463	<.0001
MEATFROZ_mean	1.09	<.0001	0.6136	<.0001
FISH_mean	-1.60	<.0001	-0.4204	<.0001
PRODUCE_mean	0.27	<.0001	0.1542	<.0001
BULK_mean	0.50	<.0001	0.5855	<.0001
SALADBAR_mean	0.50	<.0001	1.1951	<.0001
FLORAL_mean	0.04	0.5952	-0.3734	<.0001
DELI_mean	0.98	<.0001	0.1402	0.0025
CHEESE_mean	-0.66	<.0001	-0.3511	0.0317
BAKERY_mean	-0.42	<.0001	-0.5755	<.0001
GM_mean	0.39	<.0001	0.3047	<.0001
HABA_mean	0.19	<.0001	0.3665	<.0001
BEER_mean	0.15	0.0577	0.7665	<.0001
WINE_mean	-0.78	<.0001	0.6275	<.0001
SPIRITS_mean	0.29	0.073	-0.9958	<.0001
CUSTCOUN_mean	-0.68	<.0001	0.5686	<.0001

16

Fit Statistics



SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	3889.46	<.0001

What seems
to be of interest here?

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	STORE	1383624
Residual		826177
Rho = .65		

17

Prediction Error



SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

```
libname me 'C:.';

data me.pred_cs;
set me.tmp1_cs;
predms_cs = ( GROCERY_mean-pred )**2;
run;
quit;

PROC MEANS DATA=me.pred_cs
  FW=12 MEAN ;
  VAR predms_CS;
RUN;
```

18

Proc Mixed - CS

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008

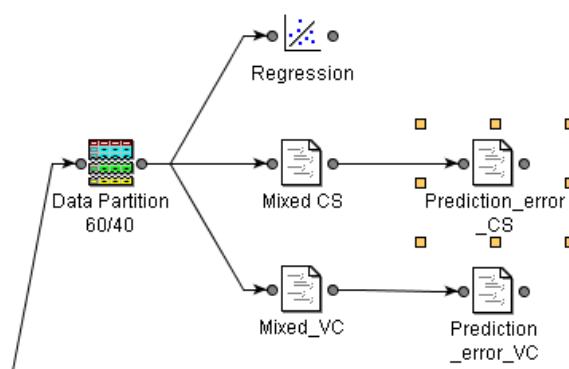
dataset	mse_vc	mse_cs	cov_cs	Prediction_VC	Prediction_CS	Percent Change
train1 60/40	1382894	826177	1383624	1340603.95	776236.49	-42.10

What seems
to be of interest here?

19

Enterprise Miner - SAS®

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008



20

Conclusion for Dominick's Finer Foods

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

- Although R^2 is large (.97) , there is correlation among the same stores sales.
- The regression model give misleading information.
- Using Compound Symmetry to model the correlation produces a model with significantly less prediction mean square error (-42%).

21

Part 2: Logistic regression in data mining

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

- When should you include quadratic predictor terms?
- When should you include interactions between predictors?
- Marginal model plots for checking model validity
- Case study

Logistic regression: why assume the logistic S-shaped curve?

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

Suppose that Y is a binary random variable (i.e., takes values 0 and 1) and that X is a single predictor variable. Then we seek a model for

$$\begin{aligned}\theta(x) &= P(Y=1 | X=x) \\ &= 1 \times P(Y=1 | X=x) + 0 \times P(Y=0 | X=x) \\ &= P(Y=1 | X=x)\end{aligned}$$

A popular choice for the S-shaped function for $\theta(x)$ is the logistic function, that is,

$$\theta(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp(-\{\beta_0 + \beta_1 x\})}$$

Solving this last equation for $\beta_0 + \beta_1 x$ gives

$$\log\left(\frac{\theta(x)}{1 - \theta(x)}\right) = \beta_0 + \beta_1 x$$

In what circumstances would you consider adding quadratic or interaction terms to a logistic regression model?

23

Logistic regression: Justification for linear and quadratic terms

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

Suppose that $X|Y=j, j=0,1$, is normally distributed with mean μ_j and variance $\sigma_j^2, j=0,1$. Then it can be shown that,

$$\log\left(\frac{\theta(x)}{1 - \theta(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$

where

$$\beta_1 = \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right), \beta_2 = \frac{1}{2} \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)$$

Thus, when the predictor variable X is normally distributed with a different variance for the two values of Y , the log odds are a quadratic function of x .

When $\sigma_1^2 = \sigma_0^2 = \sigma^2$, the log odds simplifies to

$$\log\left(\frac{\theta(x)}{1 - \theta(x)}\right) = \beta_0 + \beta_1 x$$

where

$$\beta_1 = \left(\frac{\mu_1 - \mu_0}{\sigma^2}\right)$$

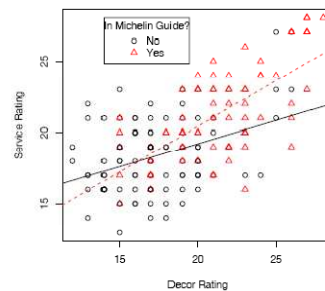
24

Logistic regression: Justification for interactions and transformations



SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

The last result can be extended to the case where we have p predictor variables which have multivariate normal conditional distributions. If the variance-covariance matrix of the predictors differs across the two groups then the log odds are a function of x_i, x_i^2 and $x_i x_j (i, j = 1, \dots, p; i \neq j)$.

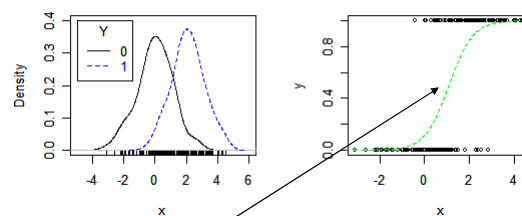


25

Simulated normal data: means=0,2; sds = 1,1



SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008

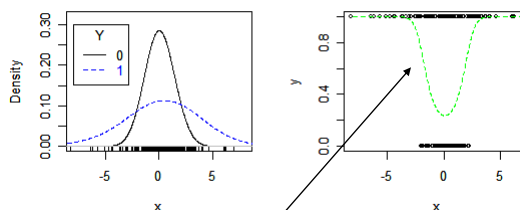


$$\theta(x) = P(Y = 1 | X = x)$$

$$\text{where } \log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x$$

26

Simulated normal data: means=0,0; sds = 1,3

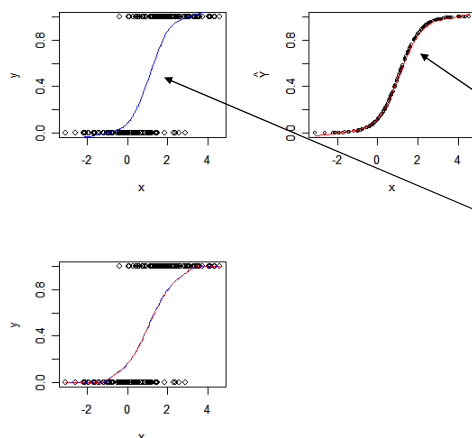


$$\theta(x) = P(Y = 1 | X = x)$$

$$\text{where } \log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$

27

Building a marginal model plot - Simulated normal data: means=0,2; sds = 1,1



$$\theta(x) = P(Y = 1 | X = x) = E(Y | X = x)$$

$$\text{where } \log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x$$

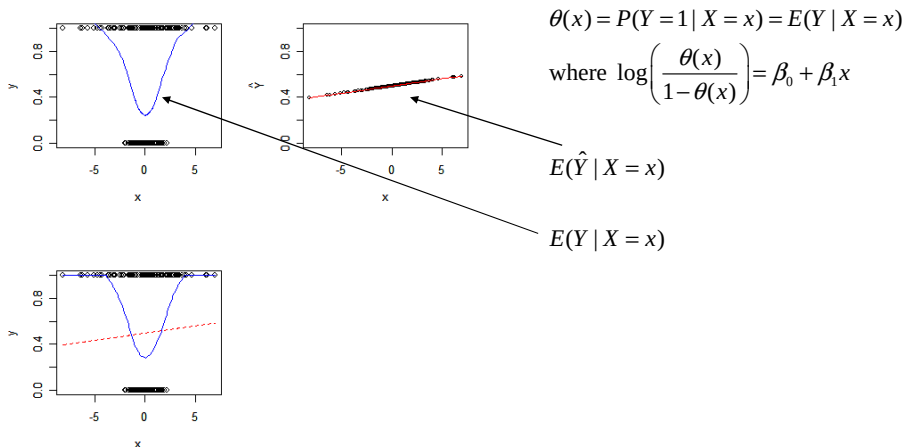
$$E(\hat{Y} | X = x)$$

$$E(Y | X = x)$$

28

Building a marginal model plot - Simulated normal data: means=0,0; sds = 1,3: Incorrect model

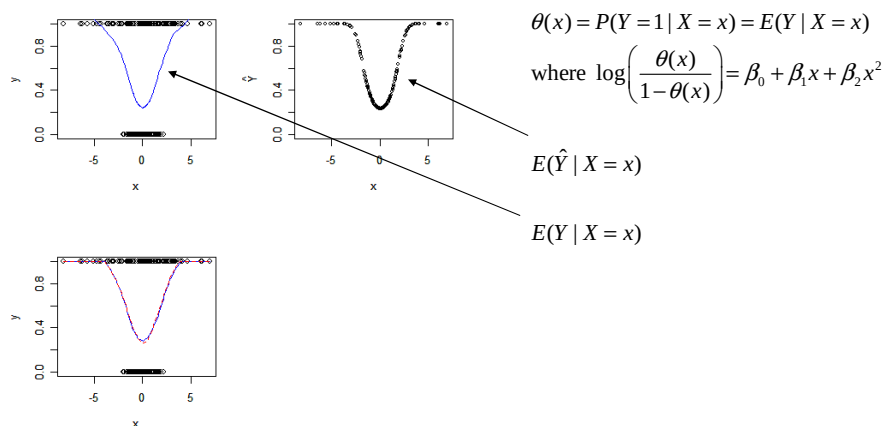
SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008



29

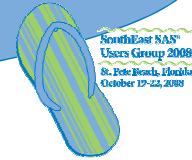
Building a marginal model plot - Simulated normal data: means=0,0; sds = 1,3: Correct model

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008



30

CASE STUDY:

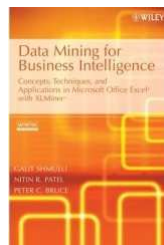


- Illustrates the importance of checking whether the current model is a valid model
- Highlights the dangers of a black box approach to modeling
- Demonstrates the value of marginal model plots in data mining

31

Direct Mail Fundraising data set:

3120 data points with 50% training & 50% validation



5.1% Response Rate
Average Donation = \$13.00
Cost To Solicit = \$0.68

AIM: Develop a model to predict

TARGET_B
1 = Donor
0 = Non-donor

from the following potential predictor variables:

ZIP: Zipcode group (zipcodes were grouped into 5 groups; only 4 are needed.)

HOMEOWNER 1 = homeowner, 0 = not a homeowner

NUMCHLD Number of children

INCOME Household income

GENDER Gender
0 = Male
1 = Female

WEALTH Wealth Rating
Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with 9 the highest wealth group and 0 the lowest. Each rating has a different meaning within each state.

HV Average Home Value in potential donor's neighborhood in \$ hundreds

ICmed Median Family Income in potential donor's neighborhood in \$ hundreds

ICavg Average Family Income in potential donor's neighborhood in hundreds

IC15 Percent earning less than 15K in potential donor's neighborhood

NUMPROM Lifetime number of promotions received to date

RAINTALL Dollar amount of lifetime gifts to date

MAXRAMNT Dollar amount of largest gift to date

LASTGIFT Dollar amount of most recent gift

TOTALMONTHS Number of months from last donation to July 1998

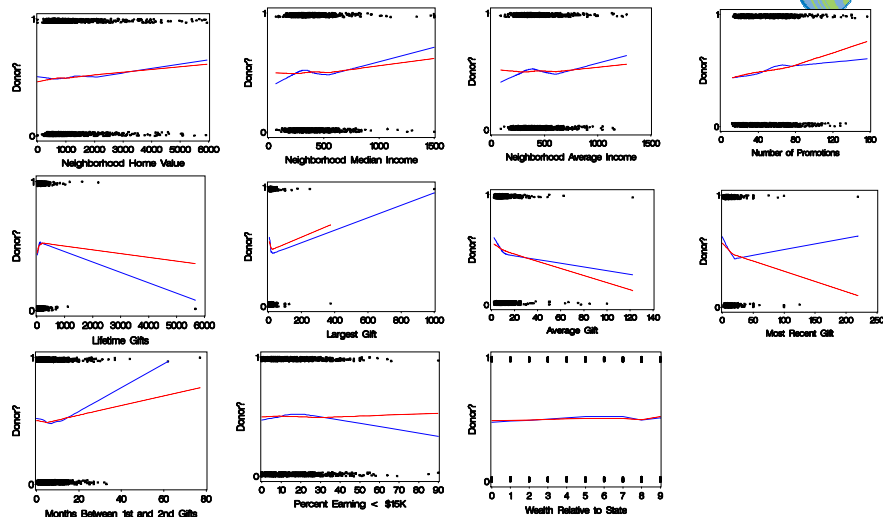
TIMELAG Number of months between first and second gift

AVGGIFT Average dollar amount of gifts to date

32

Marginal Model Plots for the full model without transformations and extra terms (MAIN EFFECTS)

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008



33

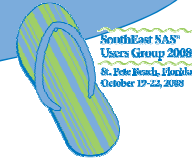
Description of the steps taken to get to the final model

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008

- Perform exploratory analysis on dataset
- Ran logistic regression
- Diagnostics indicated problems
- Used an ad-hoc Box-Cox procedure to get transformations
- Created a new factor of zip*wealth
- Conjectured that the assumptions of equal slopes by the transformed variables might not be valid
- Reran the logistic with transformed variables and interactions – did forward selection
- Diagnostics were much better

34

Final Variables and Their Transformations



- $(\text{Average Home Value})^{-.2}$
- $(\text{Median Family Income})^{.4}$
- $(\text{Average Family Income})$
- $\text{Sqrt}(\text{Number of Promotions})$
- $\text{Log}(\$ \text{ Amount of Lifetime Gifts})$
- $(\$ \text{ Amount of Largest Gift})^{.2}$
- $\text{Sqrt}(\text{Average Gift})$

35

Final Model

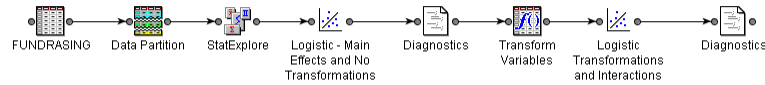


Target_B = zip_wealth
 genderdummy
 homeownerdummy
 HV_T*zip_wealth
 lcmed_T *zip_wealth
 lcavg_T*zip_wealth
 numprom_T*zip_wealth
 ramntall_T*zip_wealth
 maxramnt_T*zip_wealth
 avggft_T*genderdummy

36

Enterprise Miner SAS®

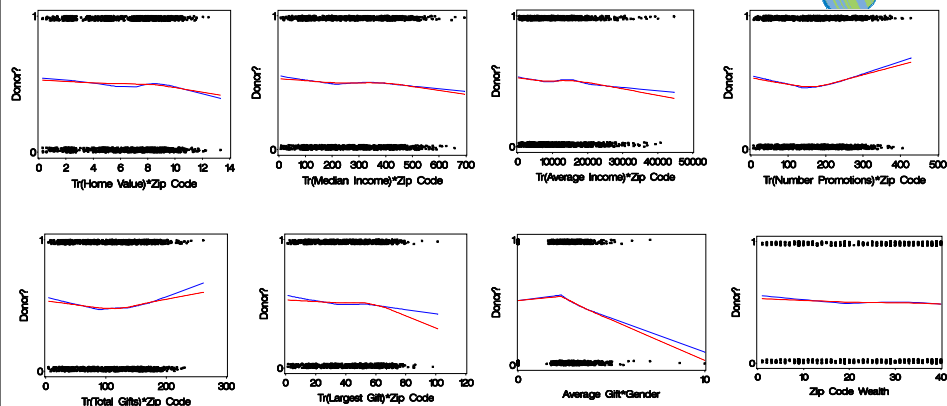
SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008



37

Marginal Model Plots for the Final model after transformations and extra terms

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-24, 2008



38

Comparison of pay-offs for original and final models – Average Profit

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008

	Training	Validation
No Transformation	0.066	0.048
Transformations	0.075	0.059

Transformations Model is 14% better for the Training Dataset and 23% better than Validation Dataset

39

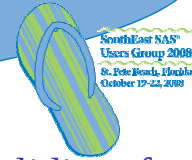
Summary

SouthEast SAS®
Users Group 2008
St. Pete Beach, Florida
October 19-22, 2008

- Illustrated the importance of checking whether the current model is a valid model
- Highlighted the dangers of a black box approach to modeling
- Demonstrated the value of marginal model plots in data mining
- Demonstrated that all this can be achieved in SAS® and SAS® Enterprise Miner

40

Conclusion



- Yes, it takes time to do checks on the validity of the model
- But not doing the checks can lead to incorrect conclusion and less than optimal models
- Statisticians working together with the data miners should be able to develop methods for doing checks on the validity of the model which take less time.

41

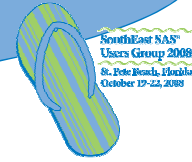
References



- Dominicks' Dataset
 - <http://research.chicagogsb.edu/marketing/databases/dominicks/index.aspx>
- “Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner” by Galit Shmueli, Nitin R. Patel, Peter C. Bruce. Wiley-Interscience (December 11, 2006)

42

About The Authors



F. Michael Speed is a Professor of Statistics at Texas A&M University

If you have any questions/comments you can contact Mike at:

Texas A&M University

3143 TAMU

College Station, TX 77843-3143

Voice(979 - 854 - 4141)

Email mspeed@stat.tamu.edu

Simon Sheather is a Professor and Head of the Department of Statistics

Contact information is the same as above

Email sheather@stat.tamu.edu