Philip Morris
Statistical Modeling Project

Time Table

To Do

① Find out who NOT to bother with.
Need to reduce the database to increase response time of system.

David Shepard Associates, Inc.
January 31, 1989

2041767097

Philip Morris

Statistical Modeling Project

Tapes due to vendor March 15 2nd - 100,000

- The immediate need is to implement a model for the March Virginia Slims mailings. Both offensive and defensive, with the offensive program having the more pressing need.

- Names for this mailing will arrive in Richmond by February 17th from three separate sources.

- The Data Set from which a model can be built consists of those individuals who:

  1. Answered a re-qualification questionnare, providing *Pre-Share-Of-Requirements* data, and answerings "a wide range of questions;"

  2. Participated in the Virginia Slims program(s) and provided Post Share-Of-Requirements data and answers to another questionnaire containing a "similar" set of questions.

Slide 1 1/31/89

2041767098

Philip Morris

Statistical Modeling Project

- The "thing(s)" the modeling will try to predict, more formally known as the "dependent variable" are:

  (a) The increase (offensive) or decrease (defensive) in Share-Of-Requirements

  (b) The increase or decrease in "Share-of-Smoker"

- The "things' the modeling will use to help in the prediction are the answers to the questions, preferably the questions asked at the time of the initial requalification questionnaire. The answers to the questions are known as the "independent" variables.

- The intitutive interpretation of the modeling effort is:

  "If we know how someone answers a questionnaire, we can estimate their relative likelihood of switching brands"

Philip Morris
Statistical Modeling Project

- *Some complications and solutions*

- The names which we'll be receiving in the future come from three sources:

  a. The Philip Morris Sweepstakes (750,000 est)
  b. Select & Save (200,000 est)
  c. FMI (500,000 est)

  The Philip Morris questionnaire contains more information (questions) than the other two sources, and less information, fewer questions, than the names on the modeling Data Set.

- Thus, the modeling effort should be restricted (initially) to using only the questions in the Data Set which correspond to questions which will be answered by the two new sources of names.

Slide 3 1/31/89

2041767100

Philip Morris
Statistical Modeling Project

- Stated another way we need one model to be applied to the Philip Morris questionnaire and another model for the other two shorter questionnaires.

- But, we really need not two models but eight models.

For Models In Which The Dependent Variable Is Change In Share Of Requirements:

Model # 1, Offensive, long set of questions
Model # 2, Defensive, long set
Model # 3, Offensive, short set
Model # 4, Defensive, short set

For Models In Which The Dependent Variable Is Change In Share Of Smoker

Model # 5, Offensive, long set
Model # 6, Defensive, long set
Model # 7, Offensive, Short set
Model # 8, Defensive, short set

Slide 4 1/31/89

2041767101

Philip Morris
Statistical Modeling Project

- For those who care (and who'll admit to not caring):
  - Share of Requirements is a continuous variable, it can take on any value from zero to 100. As such, the proper *modeling procedure is Ordinary Multiple Regression Analysis*
  - Share of Smoker is a "Yes/No" variable, someone either did switch or did not switch. The proper modeling *procedures include: Discriminant Analysis and Logistic Regression Analysis.*
- All appropriate techniques will be used

Philip Morris
Statistical Modeling Project

- All Modeling Output " The Model" looks more or less the same

- The output of a model is an equation:

$$Y = A + B1X1 + B2*X2 + B3*X3 \ldots\ldots BN*XN$$

Where in our case Y may represent either:
a. Change in Share Of Requirement
b. Change In Share Of Smoker

The X's represent the independent variables, or the questions

The B's represent the "weights" the model assigns to each variable

And, A is a constant

Philip Morris
Statistical Modeling Project

- o Other Discussion Items
- o Delivery Dates:
  - Flat Files To DSA
  - Delivery Of Statistical Models
    -- New York/Richmond/Agency Review
  - Record Layouts Of 3 Input Files
  - Programming The Name Scoring Model
  - Actual Name Scoring in Richmond
  - Names Pulled
- o Permanent Record Of Name Selection Process On Richmond Database

Philip Morris
Statistical Modeling Project

- All Modeling Output " The Model" looks more or less the same

- The output of a model is an equation:

  Y = A + B1X1 + B2*X2 +B3*X3..... BN*XN

  Where in our case Y may represent either:
  a. Change in Share Of Requirement
  b. Change In Share Of Smoker

  The X's represent the independent variables, or the questions

  The B's represent the "weights" the model assigns to each variable

  And, A is a constant

Virginia Slims

450,000 17th
300,000 27th

	A-Model SWEEPSTAKES	B-Model SELECT AND SAVE	C-Model FMI	TOTAL
GROSS NAMES	750,000	200,000	500,000	1,450,000
DUPES ON PM FILE	33.00%	33.00%	33.00%	
NET NAMES	502,500	134,000	335,000	971,500
PERCENT FEMALE	50.00%	40.00%	40.00%	
FEMALES	251,250	53,600	134,000	438,850
SMOKE PM BRANDS	60.00%	50.00%	50.00%	
AVAILABLE TO MODEL	100,500	26,800	67,000	194,300

DECILE ANALYSIS

	SWEEPSTAKES	S&S	FMI	TOTAL
DECILE 1	10,050	2,680	6,700	19,430
DECILE 2	10,050	2,680	6,700	19,430
DECILE 3	10,050	2,680	6,700	19,430
DECILE 4	10,050	2,680	6,700	19,430
DECILE 5	10,050	2,680	6,700	19,430
DECILE 6	10,050	2,680	6,700	19,430
DECILE 7	10,050	2,680	6,700	19,430
DECILE 8	10,050	2,680	6,700	19,430
DECILE 9	10,050	2,680	6,700	19,430
DECILE 10	10,050	2,680	6,700	19,430
	100,500	26,800	67,000	194,300

The long model (A) will be scored & decided differently from short (B-C) Model.

POSSIBLE SAMPLE DESIGN

	SWEEPSTAKES	S&S	FMI	TOTAL
DECILE 1	10,050	2,680	6,700	19,430
DECILE 2	10,050	2,680	6,700	19,430
DECILE 3	10,050	2,680	6,700	19,430
DECILE 4	10,050	2,680	6,700	19,430
DECILE 5	2,069	552	1,379	4,000
DECILE 6	2,069	552	1,379	4,000
DECILE 7	2,069	552	1,379	4,000
DECILE 8	2,069	552	1,379	4,000
DECILE 9	2,069	552	1,379	4,000
DECILE 10	2,069	552	1,379	4,000
	52,614	14,030	35,076	101,720

Must have the holdout sample numbers added here.

all in every category

(20% of all people in each of these)

2069 / 10,050 | 2069.00 — 2.

552 / 2680 | 552.00 — 6

20417671O6

Virginia Slims

a = (constant)

SCORING AN INDIVIDUAL

		POSSIBLE ANSWERS HIGH	POSSIBLE ANSWERS LOW	PSSSIBLE SCORES HIGH	PSSSIBLE SCORES LOW
intercept	0.00117578			0.00117578	0.0011758
Variable 1					
by pack	0.00359176	1	0	0.00359176	0.00000000
carton	-0.00179996	0	1	0.00000000	-0.00179996
both	0.00000000			0.00000000	0.00000000
Variable 2					
buy diff brand	0.00215931	1	0	0.00215931	0.00000000
buy same type	0.00000000	0	0	0.00000000	0.00000000
Variable 3					
strong d.a.	-0.04922090	0	1	0.00000000	-0.04922090
strongly agree	0.00000000	0	0	0.00000000	0.00000000
Variable 4					
disagree	-0.00033353	0	1	0.00000000	-0.00033353
agree	0.00000000	0	0	0.00000000	0.00000000
Variable 5					
21+	-0.00448282	0	1	0.00000000	-0.00448282
0-20	0.00000000	0	0	0.00000000	0.00000000
Variable 6					
years smkd = 1	0.55326560	1	0	0.55326560	0.00000000
years smkd = 3+	-0.00370587	0	1	0.00000000	-0.00370587
years smkd = 2	0.00000000	0	0	0.00000000	0.00000000
AN INDIVIDUAL'S SCORE				0.56019245	-0.0583673

Coupons

possible values of B

$$y = a + b_1X_1 + b_2X_2 + b_3X_3 \cdots b_nX_n$$

$$y = .00117578 + .00359176(X_1) + .0021593(X_2) \cdots$$

2041767107

TYPICAL RESULTS DECILE	PERCENT	IF AVG= INDEX	3 OPM	4 OPM	5 OPM	6 OPM	7 OPM	8 OPM	10 OPM
1	13.7%	238	7.1	9.5	11.9	14.3	16.6	19.0	23.8
2	9.3%	161	4.8	6.4	8.1	9.7	11.3	12.9	16.1
3	7.7%	134	4.0	5.3	6.7	8.0	9.3	10.7	13.4
4	6.5%	112	3.4	4.5	5.6	6.7	7.8	9.0	11.2
5	5.4%	94	2.8	3.8	4.7	5.6	6.6	7.5	9.4
6	4.3%	75	2.2	3.0	3.7	4.5	5.2	6.0	7.5
7	3.5%	60	1.8	2.4	3.0	3.6	4.2	4.8	6.0
8	2.9%	50	1.5	2.0	2.5	3.0	3.5	4.0	5.0
9	2.4%	42	1.3	1.7	2.1	2.5	3.0	3.4	4.2
10	2.0%	34	1.0	1.4	1.7	2.1	2.4	2.8	3.4
	5.8%	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90

RATIO OF BEST TO WORST DECILE