

Deploying Predictive Models in the Newspaper Industry

April Wilson & Kuao-Tien Chen

July 31, 2007

You Need to Have a New Direction for Your Paper



Can You Tell Me Who Was the Actress for this movie?

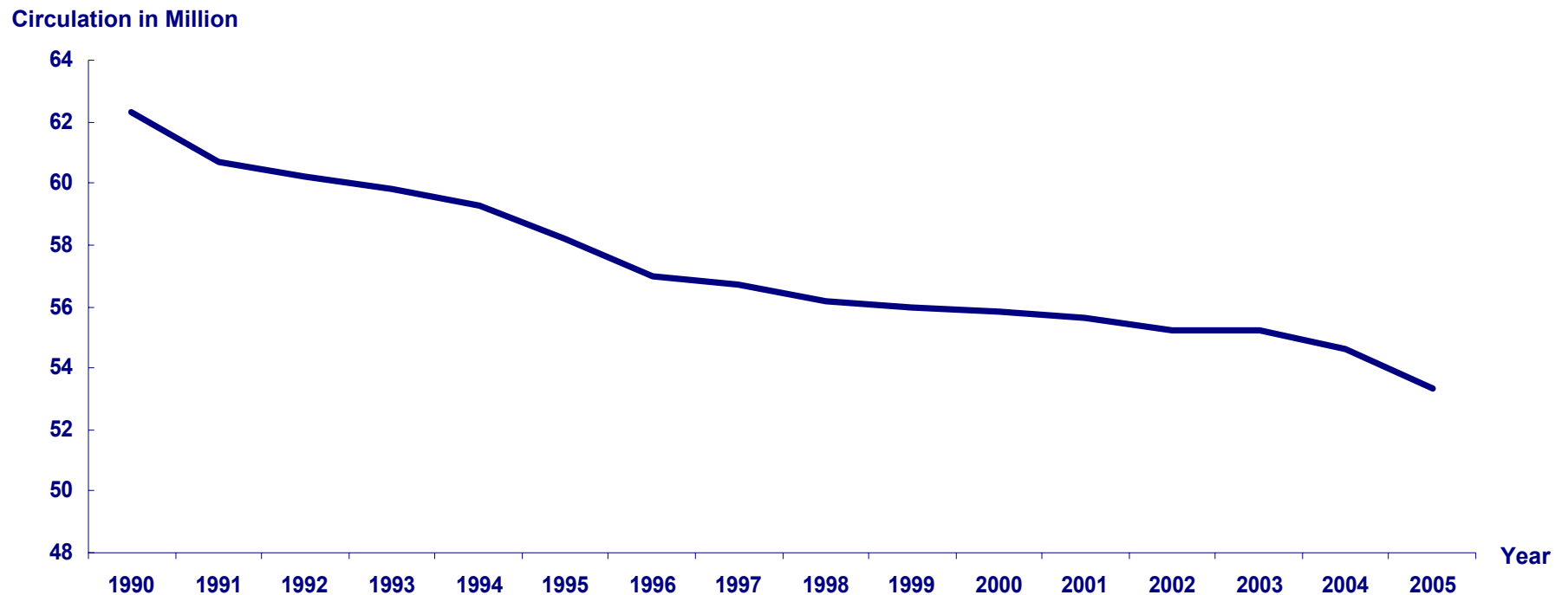


Doris Day



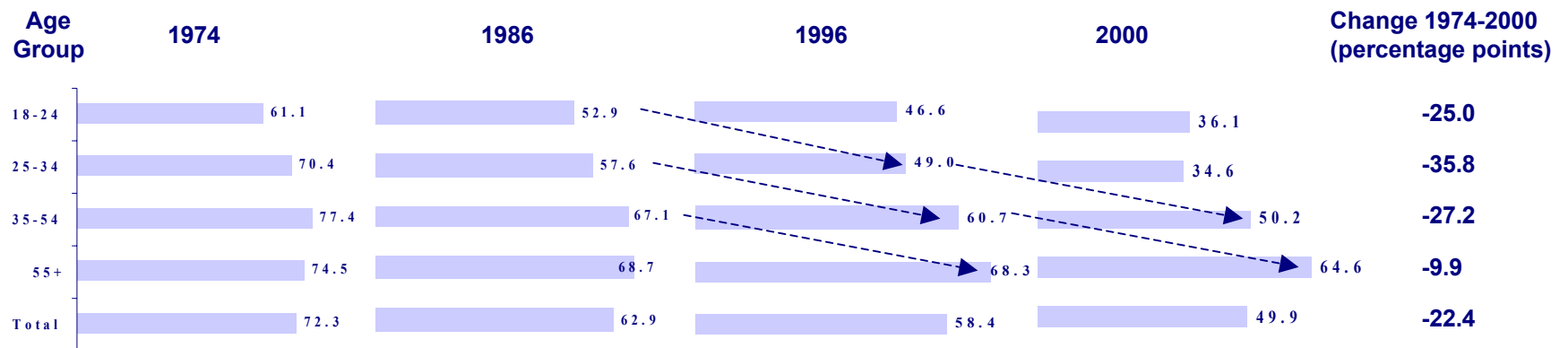
Claudette Colbert

Newspaper Circulation Continues to Decline



From 1990 to 2005, industry circulation of total weekday was down 14.4%.

Daily Newspaper Readership Percentage



The percentage of Americans reading daily newspapers has dropped from 72.3% in 1974 to 49.9% in 2000.

Paradigm Shift in The Newspaper Industry

Old World	New World
Editors select news	Consumers and aggregators select news
Communities geographically defined	Communities virtually defined
One daily source of information	Multiple daily sources of information from multiple channels
One-way advertising	Two-way system of matching buyers' wants with sellers' offerings
Consumers browse lists for information	Consumers expect precise matches to their search criteria
Advertising revenue determines success	Items sold determine success

Evolution of Media Availability Among the Generation

	Traditionalist Born Before 1946	Baby Boom Born 1946 - 1964	Generation X Born 1965 - 1976	Generation Y Born 1977 - 1994
Newspaper	√	√	√	√
Magazines	√	√	√	√
Broadcast Radio	√	√	√	√
Broadcast Television	•	√	√	√
Transistor Radio	•	√	√	√
8-track Tapes	•	•	√	√
Cassette Tapes	•	•	√	√
Walkman Radio	•	•	√	√
Video Games	•	•	√	√
VCRs	•	•	√	√
Cable Television	•	•	√	√
Personal Computers	•	•	•	√
Satellite Television	•	•	•	√
Internet	•	•	•	√
Cell Phone	•	•	•	√
Online News	•	•	•	√
DVD Players	•	•	•	•
Satellite Radio	•	•	•	•
MPs Players	•	•	•	•
TiVo	•	•	•	•
iPod Video Player	•	•	•	•

Key:

√ The technology was "always there"

• The technology was invented during this generation time

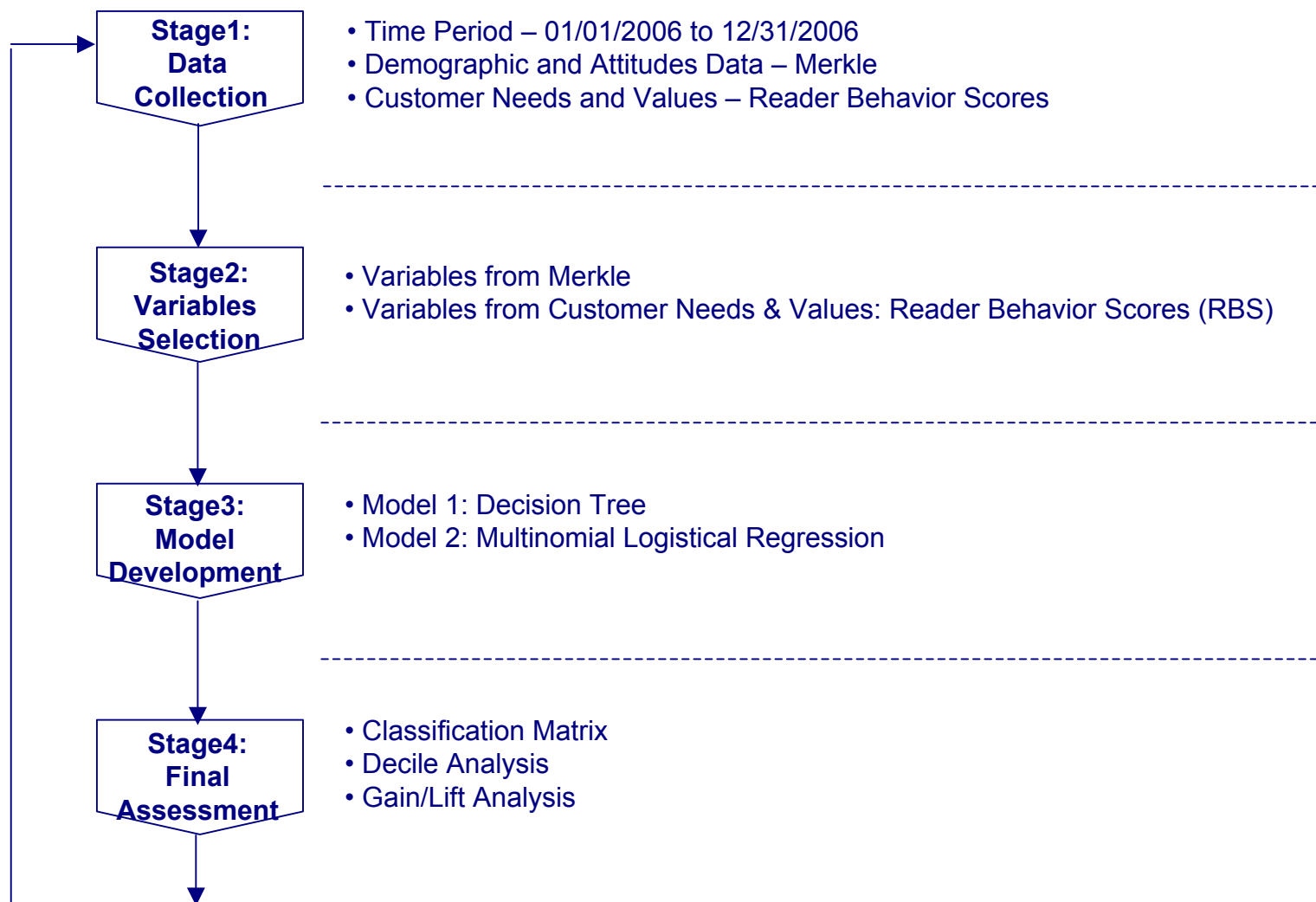
Major Challenges for The Newspaper Industry

- Demographic Factors
- Technological Innovation
- Competitive intensity
- Retailer Shift
- Changing Ownership Model

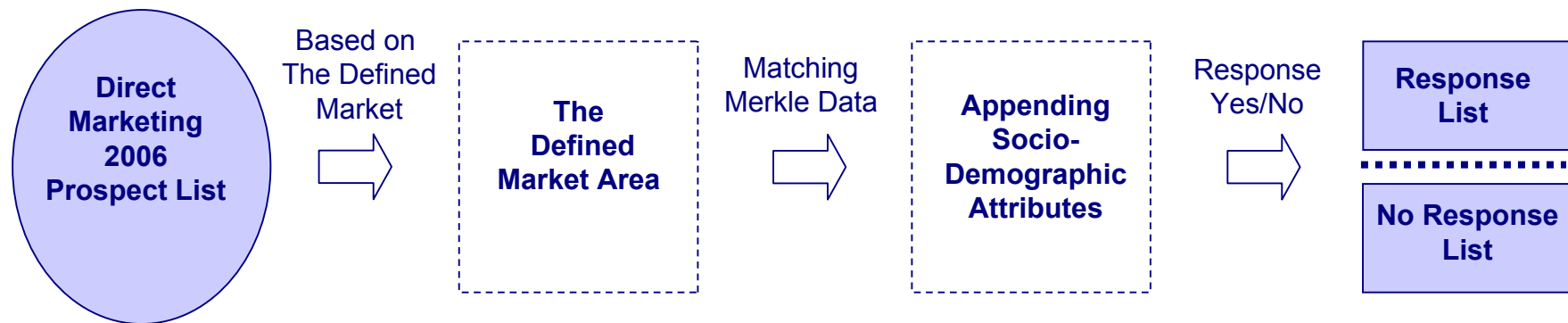
Business Objectives

- The major purpose for building the predictive model is to support the sales channel - Direct Mail to identify the best possible prospects and increase sales volumes.
- Another goal for building the predictive model is to provide the customer insights to help the company to set up the “right” offer.

Methodology



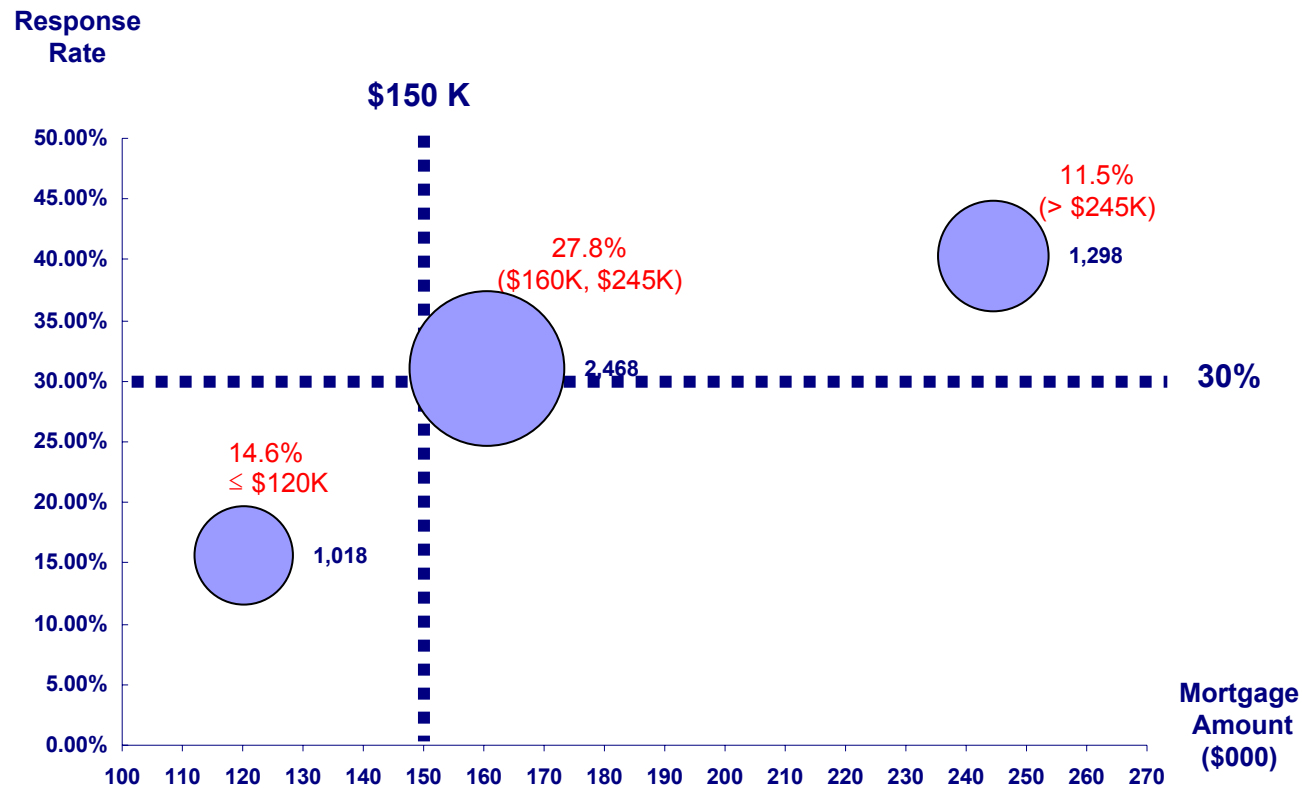
Data Sources



Consumer Profile

Disconnected to Paper	Connected to Paper	
Disconnected / Heavy Readers 22% Heavier than average newspaper readers who have a below average connection to the newspaper for the given consumer factor.	Connected / Heavy Readers 37% Heavier than average newspaper readers who have an above average connection to the newspaper for the given consumer factor.	Heavy Reader
Disconnected / Light Readers 22% Lighter than average newspaper readers who have a below average connection to the newspaper for the given consumer factor.	Connected / Light Readers 19% Lighter than average newspaper readers who have an above average connection to the newspaper for the given consumer factor.	Reader Behavior Score
		Light Reader

House Price and Newspaper Readership



The higher the house price is, the higher the response rate will be.

Reader Age and Newspaper Readership

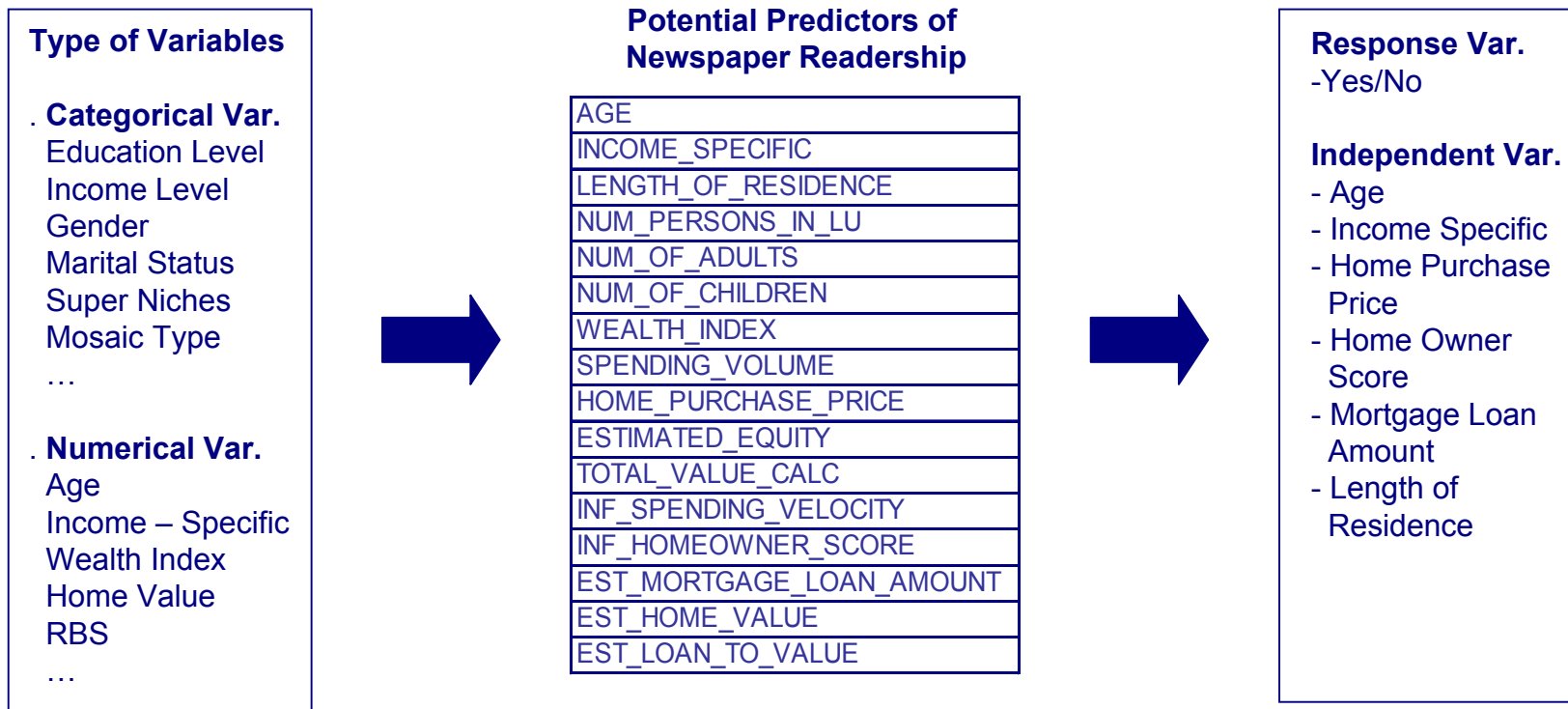
Age Range	% of Total	Reading Habits
Ages 18 to 29	3.00%	23%
Ages 30 to 49	38.94%	39%
Ages 50 to 64	32.14%	52%
Ages 65 and up	25.92%	60%

Based on the response results of 2006 Direct Mail Campaign, age is significantly related to the response rate.

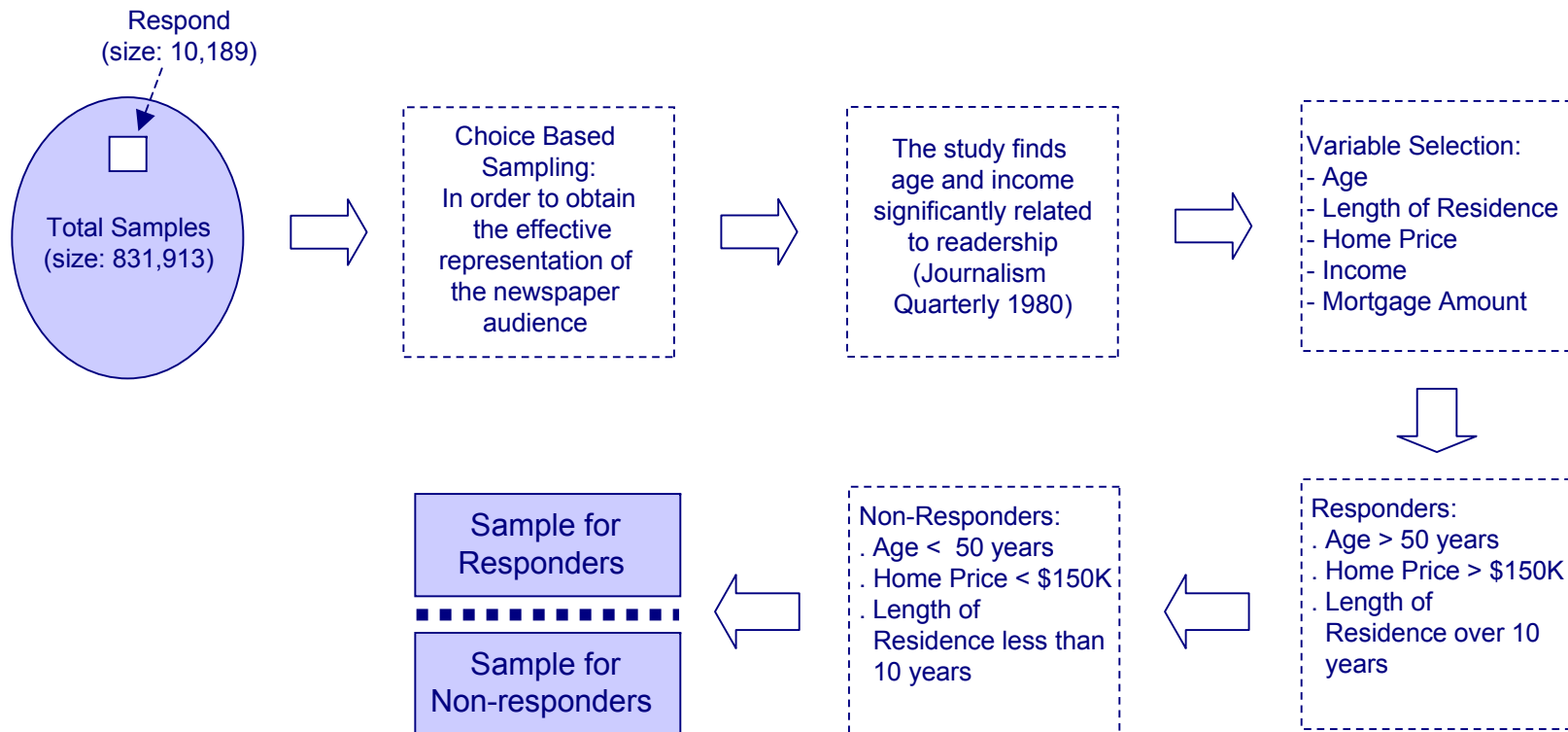
Comparing Two Means Between Response and No Response Groups

Variable	$\mu_1 - \mu_2$ (Res. vs. No Res.)	Standard Deviation	Z Value	Decision Rule (Z = 1.96)
AGE	3.27	0.19	17.18	Very Good
INCOME_SPECIFIC	0.83	0.10	8.03	Very Good
LENGTH_OF_RESIDENCE	2.05	0.11	18.41	Very Good
NUM_PERSONS_IN_LU	0.11	0.02	5.19	
NUM_OF_ADULTS	0.06	0.01	5.43	
NUM_OF_CHILDREN	0.05	0.02	2.93	
WEALTH_INDEX	6.59	1.70	3.87	
SPENDING_VOLUME	(8.38)	2.12	(3.95)	
SPENDING_VELOCITY	(9.15)	4.72	(1.94)	
NUM_OF_REGISTERED_CARS	(0.02)	0.01	(1.29)	
HOME_PURCHASE_PRICE	23295.74	2499.50	9.32	Very Good
YEAR_HOME_BUILT	(0.11)	0.25	(0.44)	
ESTIMATED_EQUITY	11833.33	1310.72	9.03	Very Good
MORTGAGE_TERM_MONTHS	(1.82)	1.46	(1.25)	
TOTAL_VALUE_CALC	21542.58	1842.29	11.69	Very Good
INF_SPENDING_VELOCITY	(1.38)	0.06	(22.08)	Very Good
INF_SPENDING_VOLUME	(0.01)	0.06	(0.18)	
INF_HOMEOWNER_SCORE	44.70	2.96	15.10	Very Good
EST_MORTGAGE_LOAN_AMOUNT	9637.23	743.68	12.96	Very Good
EST_HOME_VALUE	25130.78	1373.00	18.30	Very Good
EST_LOAN_TO_VALUE	(0.62)	0.05	(12.24)	Very Good

Variables Selection



Sampling Process



Logistical Regression Analysis

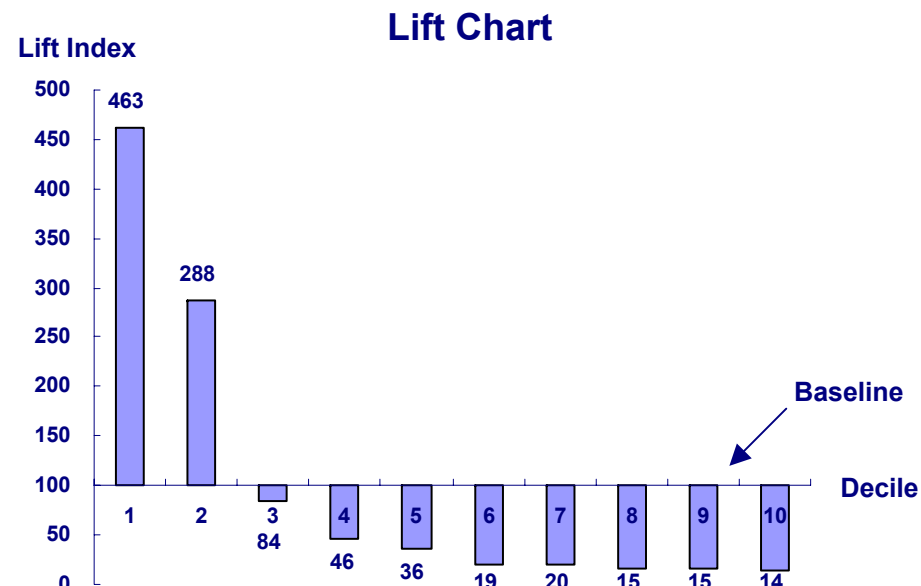
Predictive Values

Actual Values	Predictive Values		
	No Response	Response	Total
	No Response	Response	Total
	9,083 97.06%	275 2.94%	9,358 100%
	722 27.97%	1,859 72.03%	2,581 100%
	9,805 82.13%	2,134 17.87%	11,939 100%

Based on the logistical regression analysis results, the classification matrix shows that 72% of observations for this sample is assigned to the “right” category.

Response Decile Analysis for Logistical Model

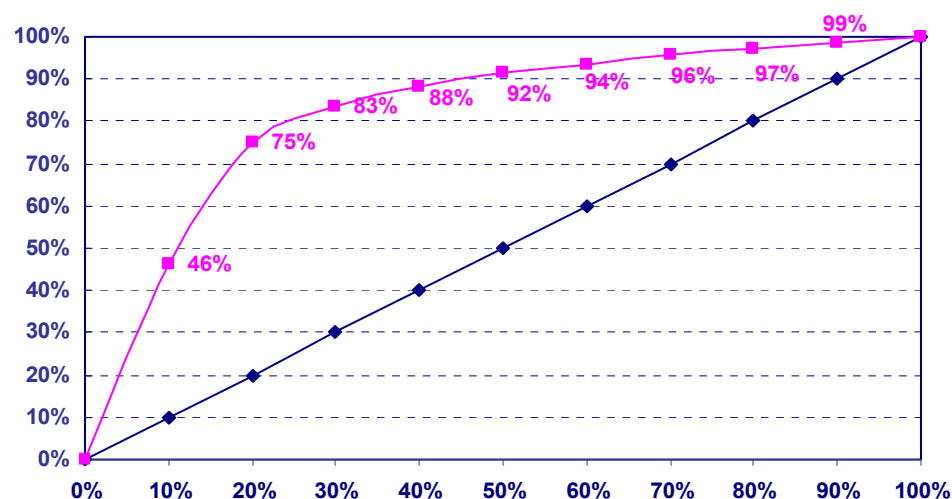
Decile	Number of Individuals	Number of Responders	Decile Response Rate	Cumulative Response Rate	Lift	Cum Lift
1	1,193	1,193	100.00%	100.00%	463	463
2	1,194	743	62.23%	81.11%	288	375
3	1,194	217	18.17%	60.12%	84	278
4	1,194	120	10.05%	47.60%	46	220
5	1,194	93	7.79%	39.64%	36	183
6	1,194	50	4.19%	33.73%	19	156
7	1,194	51	4.27%	29.52%	20	137
8	1,194	40	3.35%	26.25%	15	121
9	1,194	39	3.27%	23.69%	15	110
10	1,194	35	2.93%	21.62%	14	100
Total	11,939	2,581	21.62%			



Based on the model, top 3 deciles can generate 2.78 times the total number of responders found by randomly soliciting 30%-of-file without the model.

Cumulative Gains Chart for Logistical Model

Decile	Number of Individuals	Number of Responders	Cumulative Gains Without the Model	Cumulative Gains With the Model
1	1,193	1,193	10%	46%
2	1,194	743	20%	75%
3	1,194	217	30%	83%
4	1,194	120	40%	88%
5	1,194	93	50%	92%
6	1,194	50	60%	94%
7	1,194	51	70%	96%
8	1,194	40	80%	97%
9	1,194	39	90%	99%
10	1,194	35	100%	100%
Total	11,939	2,581		



83% of responders are identified in top 30% of those samples.

Results

- The statistical results for the logistical regression analysis show us that the model has the capability to define a prospect on the mail list who will respond the direct mail at the high probability (more than 70%).
- The Lift/Gain charts indicate that the model is robust and consistent. Without the model, if we mail 30% of the file, we will get 30% of the potential responders. If we use the model and mail the same percentage, we capture over 83% of responders. It means that with the model, we can use 30% of marketing cost to generate 83% of responders.

Responder Profile

Based on 2006 direct mail campaign, the typical DM responder is more likely to

- Be in the top 30% decile of Homeowner Score,
- Have household income > \$75,000,
- Have more than 9 years length of residence,
- Have house price over \$150,000,
- Belong to the single home segment,
- Have more than average number of real property trades (estimated mortgage loan amount),
- Be an older age (over 50).

Q & A