The Art and Science of Constructing Powerful Predictors

How to exploit your customer data

December 7, 1999

NCDM Orlando '99 Conference and Exhibition

Perry D. Drake, Vice President, Drake Business Services, Inc.
& Rhonda Knehans Drake, President, Drake Business Services, Inc.

Presenters

Perry D. Drake

Vice President, Drake Business Services, Inc.

- **Independent database marketing consultant since 1998.**
- **Also since 1998, teach statistical and database modeling and analysis techniques to future direct marketers at New York University in their Direct Marketing Master's Degree Program.**
- **From 1987 to 1998 worked at the The Reader's Digest Association:**
  - **Director, Marketing Services**
  - **Association Marketing Director, Magazine Circulation**
  - **Senior Quantitative Analyst**
- **Master of Science in Applied Statistics from the University of Iowa and a Bachelor or Science in Economics from the University of Missouri.**

Presenters (Continued)

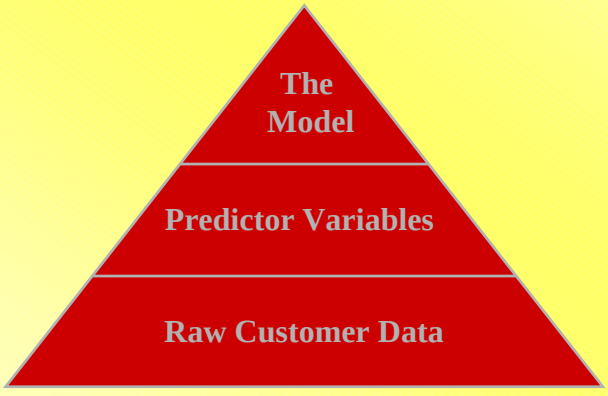
Rhonda Knehans Drake

President, Drake Business Services, Inc.

- **Formed database marketing consulting business in 1997.**
- **From 1987 to 1997 worked in various information management roles:**
  - **Director of Customer Database and List Management, Reader's Digest Young Families**
  - **Senior Account Executive, Information Resources, Inc.**
  - **Manager, Market Planning and Analysis, Video Continuity, Columbia House**
- **Master of Science in Applied Statistics from the University of Iowa and a Bachelor or Science in Economics from the University of Missouri.**

Why This Topic

To highlight the importance of the analytical step to creating predictors to ensure success in the modeling/analysis step.



Objective

Upon leaving this presentation you will have a solid understanding of:

- **The importance of understanding the data residing on your customer database before beginning any analysis.**
- **How to become intimately familiar with your customer file.**
- **The various techniques used to manipulate and exploit customer data.**

Getting to Know Your Customer Data

Before embarking on any analysis project involving a customer database, a solid understanding of the customer file and the data contained on it is absolutely necessary.

Without fully understanding your customer data, how can you possibly exploit it to its fullest predictive potential!

Getting to Know Your Customer Data (Continued)

To help you in understanding your customer data:

- **Meet with the staff/vendor responsible for creating and maintaining the marketing data.**
- **Meet with the individuals responsible for maintaining and applying field updating rules.**

Getting to Know Your Customer Data (Continued)

Full comprehension of the data residing on the database is crucial for any analysis. Questions to consider include:

- **Does a partial payment for a product order update the "Total Products Paid" field?**
- **If so, is this consistent across all product lines?**
- **If a customer orders, pays and then returns product, is the "Total Products Paid" field properly decremented?**
- **What is the method for recording a "claims paid"?**

Data Manipulation

When preparing to build a target model or segment a customer file, there are numerous way to assess and maximize the predictive strength of various data elements. The techniques we will discuss include:

- **Univariate analysis**
- **Cross-tabulations**
- **Manipulation of recency, frequency & monetary values**
- **Creation of logic variables**
- **Development of ratio variables**
- **Creation of longitudinal variables**
- **Time alignment of key events**
- **Creation of "in-house" zip level data elements**

Methodology Assumptions

Before beginning our discussion regarding the creation of variables, we will assume:

- **You conducted a test to a random sample of names from your database.**
- **You saved all customer information at point-in-time of the promotion.**
- **Your goal is to determine the most powerful customer attributes for modeling customer behavior.**

Univariate Analysis

Univariate analysis examines one variable at a time to determine its strength in terms of helping predict the customer behavior of concern (order, payment, cancel, etc.).

- **The most commonly used form of analysis**
- **Too often the only form of analysis employed by many direct marketers**

Univariate Analysis (Continued)

Assume a direct marketer of books, music and video products test promotes a new music package "Rock of the 80s" to a sample of 10,000 customers taken randomly from the customer database.

We will consider and discuss two univariate views of the customer data residing on this sample:

- **Total Number of Orders**
- **Total Number of Promotions**

Univariate Analysis (Continued)

TOTAL NUMBER OF ORDERS EVER (ALL PRODUCT LINES)

Test promotion: "Rock of the 80s"
Sample Size: 10,000
Overall Response Rate: 2.50%

The response rate for each category divided by the response rate of the total sample.

Total Number of Orders Ever (all Prod. lines)	Number	% of Sample	Number of Orders	Response Rate	Index to Total
1-5	3,312	33.12%	62	1.87%	75
6-10	3,074	30.74%	68	2.21%	88
11-15	2,205	22.05%	64	2.90%	116
16 plus	1,409	14.09%	56	3.97%	159
Total	10,000	100.00%	250	2.50%	100

The various categories for the variable.

The number of names falling into each category.

The number of orders falling into each category.

The response rate for each category which is calculated as the number of orders divided by the number of names in each category.

Univariate Analysis (Continued)

TOTAL NUMBER OF ORDERS EVER (ALL PRODUCT LINES)

Test promotion: "Rock of the 80s"
Sample Size: 10,000
Overall Response Rate: 2.50%

Total Number of Orders Ever (all Prod. lines)	Number	% of Sample	Number of Orders	Response Rate	Index to Total
1-5	3,312	33.12%	62	1.87%	75
6-10	3,074	30.74%	68	2.21%	88
11-15	2,205	22.05%	64	2.90%	116
16 plus	1,409	14.09%	56	3.97%	159
Total	10,000	100.00%	250	2.50%	100

Using order information alone, given a break-even response rate of 3.00%, which names can the product manager profitably promote?

Univariate Analysis (Continued)

TOTAL NUMBER OF ORDERS EVER (ALL PRODUCT LINES)

Test promotion: "Rock of the 80s"
Sample Size: 10,000
Overall Response Rate: 2.50%

Total Number of Orders Ever (all Prod. lines)	Number	% of Sample	Number of Orders	Response Rate	Index to Total
1-5	3,312	33.12%	62	1.87%	75
6-10	3,074	30.74%	68	2.21%	88
11-15	2,205	22.05%	64	2.90%	116
16 plus	1,409	14.09%	56	3.97%	159
Total	10,000	100.00%	250	2.50%	100

Using order information alone, the product manager can promote anyone on the database that has placed 16 or more orders in the past.

Univariate Analysis (Continued)

Let's now consider a univariate tabulation of "Total Number of Promotions Ever."

Univariate Analysis (Continued)

TOTAL NUMBER OF PROMOTIONS EVER (ALL PROD. LINES)

Test promotion: "Rock of the 80s"
Sample Size: 10,000
Overall Response Rate: 2.50%

Total Number Promotions Ever (all Prod. lines)	Number	% of Sample	Number of Orders	Response Rate	Index to Total
1-5	768	7.68%	16	2.08%	83
6-10	2,544	25.44%	57	2.24%	90
11-20	3,563	35.63%	108	3.03%	121
21 plus	3,125	31.25%	69	2.21%	88
Total	10,000	100.00%	250	2.50%	100

Notice that this variable does not show a nicely increasing or decreasing response rate as the number of promotions increases as was the case with the "orders" variable.

Univariate Analysis (Continued)

TOTAL NUMBER OF PROMOTIONS EVER (ALL PROD. LINES)

Test promotion: "Rock of the 80s"
Sample Size: 10,000
Overall Response Rate: 2.50%

Total Number Promotions Ever (all Prod. lines)	Number	% of Sample	Number of Orders	Response Rate	Index to Total
1-5	768	7.68%	16	2.08%	83
6-10	2,544	25.44%	57	2.24%	90
11-20	3,563	35.63%	108	3.03%	121
21 plus	3,125	31.25%	69	2.21%	88
Total	10,000	100.00%	250	2.50%	100

BE CAREFUL!
Promotion data alone will be an unstable predictor of customers likely to respond and not respond to a particular offer. Why?

Univariate Analysis (Continued)

Because a customer has been promoted more than another, it does not mean they are a better customer.

- **Those with more promotions have obviously been on file longer but some may have stopped ordering while others continue to order and are, therefore, better customers.**
- **Those with few promotions are a mix of "new-to-file" customers and "older" customers that simply stopped ordering soon after they came on file.**

Cross-Tabulations

Cross-tabulations are a means of viewing two or more data elements in combination.

- **Highlights interrelationships among variables.**
- **Strengthens relatively weak or unstable data elements.**

Cross-Tabulations (Continued)

	Total Promotions Ever:				
Total Orders Ever:	1-5	6-10	11-20	21 plus	Total
1-5	1.63% (8/491)	1.76% (17/967)	2.34% (20/856)	1.60% (16/998)	1.87% (62/3,312)
6-10	2.89% (8/277)	1.85% (14/756)	2.51% (29/1,154)	1.80% (16/887)	2.21% (68/3,074)
11-15	0.00% (0/0)	3.03% (14/462)	3.03% (29/956)	2.67% (21/787)	2.90% (64/2,205)
16 plus	0.00% (0/0)	3.34% (12/359)	5.03% (30/597)	3.53% (16/453)	3.97% (56/1,409)
Total	2.08% (16/768)	2.24% (57/2,544)	3.03% (108/3,563)	2.21% (69/3,125)	2.5% (250/10,000)

This figure represents the response rate for customers with "6-10" promotions and "1-5" orders.

The "Total" for each column represents the "univariate figures" for the variable "Total Promotions Ever."

The "Total" for each row represents the "univariate figures" for the variable "Total Orders Ever."

Cross-Tabulations (Continued)

	Total Promotions Ever:				
Total Orders Ever:	1-5	6-10	11-20	21 plus	Total
1-5	1.63% (8/491)	1.76% (17/967)	2.34% (20/856)	1.60% (16/998)	1.87% (62/3,312)
6-10	2.89% (8/277)	1.85% (14/756)	2.51% (29/1,154)	1.80% (16/887)	2.21% (68/3,074)
11-15	0.00% (0/0)	3.03% (14/462)	3.03% (29/956)	2.67% (21/787)	2.90% (64/2,205)
16 plus	0.00% (0/0)	3.34% (12/359)	5.03% (30/597)	3.53% (16/453)	3.97% (56/1,409)
Total	2.08% (16/768)	2.24% (57/2,544)	3.03% (108/3,563)	2.21% (69/3,125)	2.5% (250/10,000)

Using this cross-tabulation information, given a break-even response rate of 3.00%, which names can the product manager profitably promote?

Cross-Tabulations (Continued)

	Total Promotions Ever:				
Total Orders Ever:	1-5	6-10	11-20	21 plus	Total
1-5	1.63% (8/491)	1.76% (17/967)	2.34% (20/856)	1.60% (16/998)	1.87% (62/3,312)
6-10	2.89% (8/277)	1.85% (14/756)	2.51% (29/1,154)	1.80% (16/887)	2.21% (68/3,074)
11-15	0.00% (0/0)	3.03% (14/462)	3.03% (29/956)	2.67% (21/787)	2.90% (64/2,205)
16 plus	0.00% (0/0)	3.34% (12/359)	5.03% (30/597)	3.53% (16/453)	3.97% (56/1,409)
Total	2.08% (16/768)	2.24% (57/2,544)	3.03% (108/3,563)	2.21% (69/3,125)	2.5% (250/10,000)

By cross-tabbing the "orders" and "promotion" counter variables the product manager can profitably promote a portion of the names that have "11-15" orders.

Manipulation of Recency, Frequency & Monetary Values

Don't restrict the use of recency, frequency and monetary value data to straight univariate views. Be creative!

- **Combine these three elements**
- **Split the elements by product line**
- **Split the elements by affinity, within a product line**
- **View them by more relevant time periods**

Manipulation of Recency, Frequency & Monetary Values (Continued)

Consider the following variations on the variable "Total Number of Orders Ever":

- **Total number of orders in the past 12 months**
- **Total number of orders in the past 12-24 months**
- **Total number of cookbook orders in the past 36 months**
- **Total number of paid music orders (any affinity) in the past 24 months**
- **Total number of cancels/returns in the past 24 months**
- **Total number of cancels/returns vs total number of paids**

Manipulation of Recency, Frequency & Monetary Values (Continued)

Consider the following variations on the variable "Last Order Date":

- **Last cookbook order date**
- **Last music order data (any affinity)**
- **Elapsed time in months since last paid order**
- **Elapsed time in months since last cancel/return**
- **Elapsed time in months since last club order**

Logic Variables

When similar data elements are combined into one, it is termed a *logic variable*. A logic variable is a counter variable (but not all counter variables are logic variables).

- **A logic variable created from many similar, weaker variables can become a strong data element.**
- **The creation of logic variables also reduces the odds of the final regression model having multicollinearity.***

**** Multicollinearity, a violation of any multiple regression model, occurs when two or more predictor variables in the final model are correlated with one another.***

Logic Variables (Continued)

Assume the analyst, in determining who to promote for "Rock of the 80s," has decided to consider examining past rock music purchase data:

- **Rock and Roll Party (RRP)**
- **The Soul of Rock and Roll (TSRR)**
- **Rock Legends (RL)**
- **Easy Listening Rock (ELR)**

Logic Variables (Continued)

To create a logic based on this data, you add up the number of the past rock music packages each customer has ordered. This "Rock Music" logic variable will take on the values 0, 1, 2, 3 or 4.

After the logic is created on the test sample, the analyst will run a univariate tabulation to determine the strength of this new variable in terms of identifying responders for "Rock of the 80s."

Logic Variables (Continued)

"Rock Music" Logic Variable

Test promotion: "Rock of the 80s"
Sample Size: 10,000
Overall Response Rate: 2.50%

Rock Logic: (RRP, TSRR, RL, ELR)	Number	% of Sample	Number of Orders	Response Rate	Index to Total
Purchased 0	7,856	78.56%	118	1.50%	60
Purchased 1	945	9.45%	47	5.01%	200
Purchased 2	633	6.33%	40	6.34%	254
Purchased 3	365	3.65%	27	7.44%	298
Purchased 4	201	2.01%	18	8.87%	355
Total	10,000	100.00%	250	2.50%	100

Ratio Variables

Ratio variables are the result of dividing one continuous data element by another. The creation of ratio variables, like cross-tabulations, can turn relatively weak predictors into strong predictors.

Ratio Variables (Continued)

Ratio Variable	Measurement of
Total Products Paid for each customer divided by Total Products Ordered	Estimate of average payment rate for each customer
Total Book Products Paid divided by Total Products Paid	Strength of the book affinity as opposed to others (music, video)
Total Orders divided by Total Promotions Sent	Overall responsiveness to all promotions
Total Music Paid Orders divided by Total Music Promotions Sent	Estimate of average paid response rate for music orders

Longitudinal Variables

Longitudinal (or time-series) variables:

- **Provide a view of a particular data element for each customer across time.**
- **Are based on the premise that the best predictor of a customer's next reaction to a promotion is the customer's reactions to the most recent promotions.**
- **Can assist you in "life-stage" marketing efforts.**

Longitudinal Variables (Continued)

Longitudinal Variable	Measurement of
Responses (order, pay, cancel, non-response, etc.) to the last 3 promotions sent	Estimate of customer's action on next promotion sent
Last 4 product affinity shipments for the music product line (rock, pop, country, etc.)	Estimate of most current music interests
Limiting either of the above variables to within the past year guarantees the information will be fresh and relevant.	
Last 2 responses to the "music affinity" question on a customer questionnaire	Gauge of the next most likely affinity purchase
To develop this variable, you must save old questionnaire data after updates.	

Longitudinal Variables (Continued)

Assume the analyst created a longitudinal variable revealing each customer's last three actions to promotions sent on the test sample of 10,000 names.

Following is the resulting univariate tabulation of this longitudinal variable.

Longitudinal Variables (Continued)

Last Three Customer Actions

Test promotion: "Rock of the 80s"
Sample Size: 10,000
Overall Response Rate: 2.50%

Customer Actions to Last Three Promotions Sent (Prom-2, Prom-1, Prom)	Number	% of Sample	Number of Orders	Response Rate	Index to Total
(Pay, Pay, Pay)	356	3.56%	19	5.34%	213
(NR, Pay, Pay)	422	4.22%	16	3.79%	152
(NR, NR, Pay)	528	5.28%	18	3.41%	136
(Cancel, Pay, Pay)	427	4.27%	20	4.68%	187
(Cancel, Cancel, Pay)	229	2.29%	9	3.93%	157
……	……	……	……	……	……
(NR, NR, BD)	150	1.50%	2	1.33%	53
(NR, NR, NR)	298	2.98%	2	0.67%	27
Total	10,000	100.00%	250	2.50%	100

NR = Non-Response, BD = Bad Debt

Time Alignment of Key Events

Marketers who regularly communicate with customers due to the continuous nature of their business are well positioned to leverage the customer dynamic. Such companies include:

- **Clubs**
- **Continuities or collectible marketers**
- **Frequent buyer clubs**
- **Frequent travel services**
- **Banking and financial services**
- **Telecommunications and cable companies**

Time Alignment of Key Events
(Continued)

How to assess the customer dynamic

- **Profits depend on quickly identifying the best customers.**
- **Must distinguish between:**
  - **long term loyalty**
  - **attrition**
- **Two parameters are required to assist in identifying homogeneous groups:**
  - **critical time point**
  - **defining behavior**

Time Alignment of Key Events
(Continued)

The critical time point may be a key purchase behavior or a "life stage" event presenting a new communication opportunity. For example:

- **Initial enrollment in service**
- **Purchase of first up-sell product**
- **Purchase of home**

Time Alignment of Key Events
(Continued)

To develop predictors defining your best customers align all accounts on initial enrollment in service. The following predictors can be built as a measure of quality:

- **Days to order first product following enrollment**
- **Ratio variables measuring dollars paid to total number of promotions**
- **Counter variables representing the total number of promotions with no activity**

Time Alignment of Key Events
(Continued)

If data has been aligned on the first up-sell product purchase then consider:

- **Time to pay for first up-sell product.**
- **Acceptance (non-return, non-cancel) of first up-sell, as a measure of effectiveness in customer relationship management.**

Time Alignment of Key Events
(Continued)

Time align customer accounts based on "life-stage" events and develop predictors such as:

- **Ratio of the dollars paid on home mortgage to the total months of home ownership**
- **Cross-tabulation of household income versus total home equity**

The Creation of In-House Zip Level Data Elements

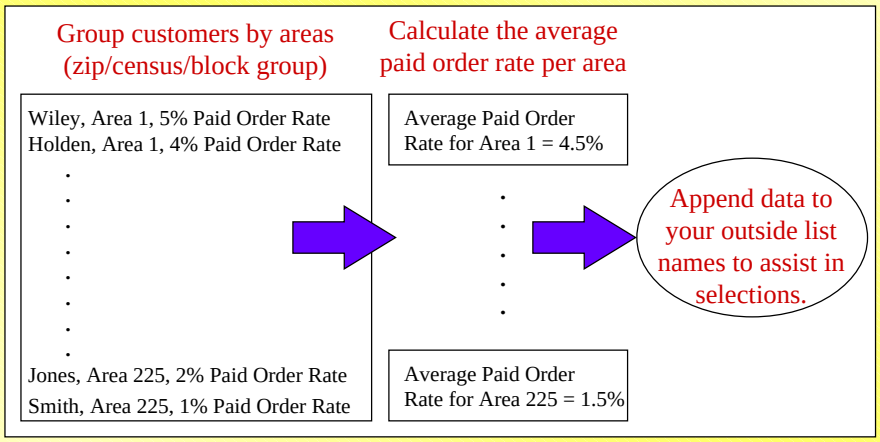
If you need to increase the response/payment rates for your outside list programs or "data poor" segments, consider this technique:

- **Drive your in-house customer data down to a zip code, census tract or block group level (depending on the size of your customer file).**
- **Use this in-house "zip" level data to help you determine which areas to promote for your particular product/ service.**
- **This technique will strengthen your selections to outside lists and "data poor" customer segments.**

The Creation of In-House Zip Level Data Elements (Continued)

Assume the analyst decided to drive customer payment information down to a zip level in order to help determine those customers living in areas most capable of affording an expensive product offering.

The Creation of In-House Zip Level Data Elements (Continued)



Summary

Preparing for Analysis

- **Have a solid understanding of the customer data**
- **Ensure the data residing on the database is being properly updated and maintained**

Beginning the Analysis

- **Clearly define you objective**
- **Creatively brainstorm how you can manipulate the customer data in order to successfully meet your goals**

The idea is to become as knowledgeable about your customer data as possible.