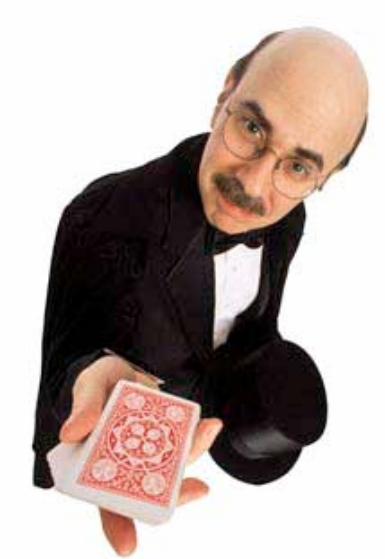
Demystifying the Modeling Process

An Introduction to Database Marketing Practices

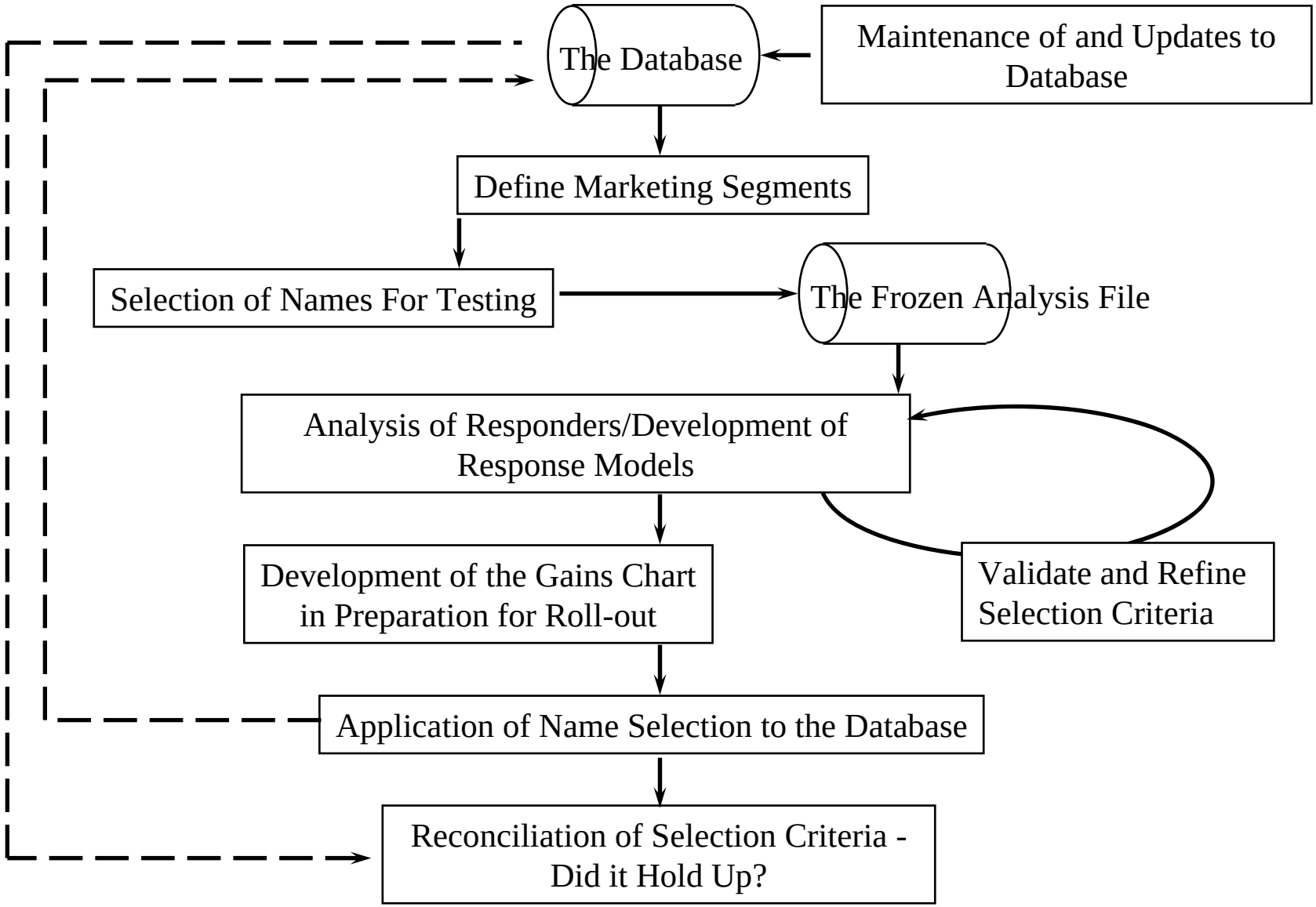


Perry D. Drake © 2005
Vice President, Drake Direct, New York, NY
Associate Professor, New York University
www.drakedirect.com

Database and Analysis Introduction

We will begin with an overview of the database analysis process flow as shown on the next slide.

Database Analysis Cycle



Course Outline

For this session we will discuss the following topics of the database analysis process flow. Due to limited time, each topic will be covered at a high level.

- Overview of the Marketing Database
- Segmenting the Customer Database
- Selecting the Test Sample
- Analyzing Results of the Test Sample
- Modeling Responders to the Test Offer
- Developing the Gains Chart

Section 1
Overview of the Marketing Database

Data Types

There are Two major types of data residing on a marketing database:

- Internal or House Data
- External or Enhancement Data

We will not be concerning ourselves with fulfillment data or databases today.

House Marketing Data

Core house marketing data is made up of promotional and purchase data and can be classified as follows:

- **R**ecency data
- **F**requency data
- **M**onetary data

This information is also known as **RFM** data and plays a key role in the segmentation and response modeling.

Enhancement Data

Enhancement data is information about a direct marketer's customers obtained from a third party. This type of data is typically purchased by a direct marketer in an attempt to supplement their own "house" customer data with new information. It is appended to the customer database for purposes of:

- Learning more about the customers
- Assisting a product manager in better selecting customers on the database for future promotions
- Assisting the editorial division in new product development

Enhancement Data (Continued)

The two main categories of external data discussed in this section are:

- Census data
- Compiled list data

Census Data (Continued)

The US Government gathers census data every 10 years, re-estimating some data, such as population growth estimates, between updates. Census data available within a geo-demographic region includes:

- Average income
- Average household size
- Average home value
- Average monthly mortgage
- Percent ethnic breakdown
- Percent married
- Percent college education
- Even such measures as average daily commuting time!

Compiled List Data

Compiled list data is individual data collected by service bureaus for the purpose of selling it to direct marketers. This data can be compiled from a variety of sources:

- Personal data listed on product warranty cards or rebate vouchers
- Questionnaire information provided in order to receive free product samples and coupons
- Vehicle registration data
- Web site registration information
- Credit report data (with restrictions)

Compiled List Data (Cont.)

Complied list data may be categorized as either demographic or psychographic (lifestyle) in nature:

Demographic:

- Income
- Age
- Gender
- Education
- Location
- Language
- Ethnicity
- Marital status
- Children
- Occupation

Psychographic:

- Hobbies
- Reading interests
- Exercising habits
- Music preferences
- Move preferences
- Attitudes and beliefs
- Lifestyle
- Life stage

Section 2

Segmenting the Customer Database

Why We Segment

The underlying premise for segmentation of your database is that not all customers residing on the database are alike and, therefore, should not be treated alike. Segmenting the customer database into sub-markets will allow you to more effectively market various products, services and offers to your customers.

The main goal of a segmentation scheme is to group customers into homogeneous groups that fit the intended objective. These groupings enable more effective and efficient marketing efforts to, communications with, and research of those customer segments.

Why We Segment (Cont.)

A customer file can have several overall segmentation schemes depending on the objective to be attained. Possible segmentation objectives are:

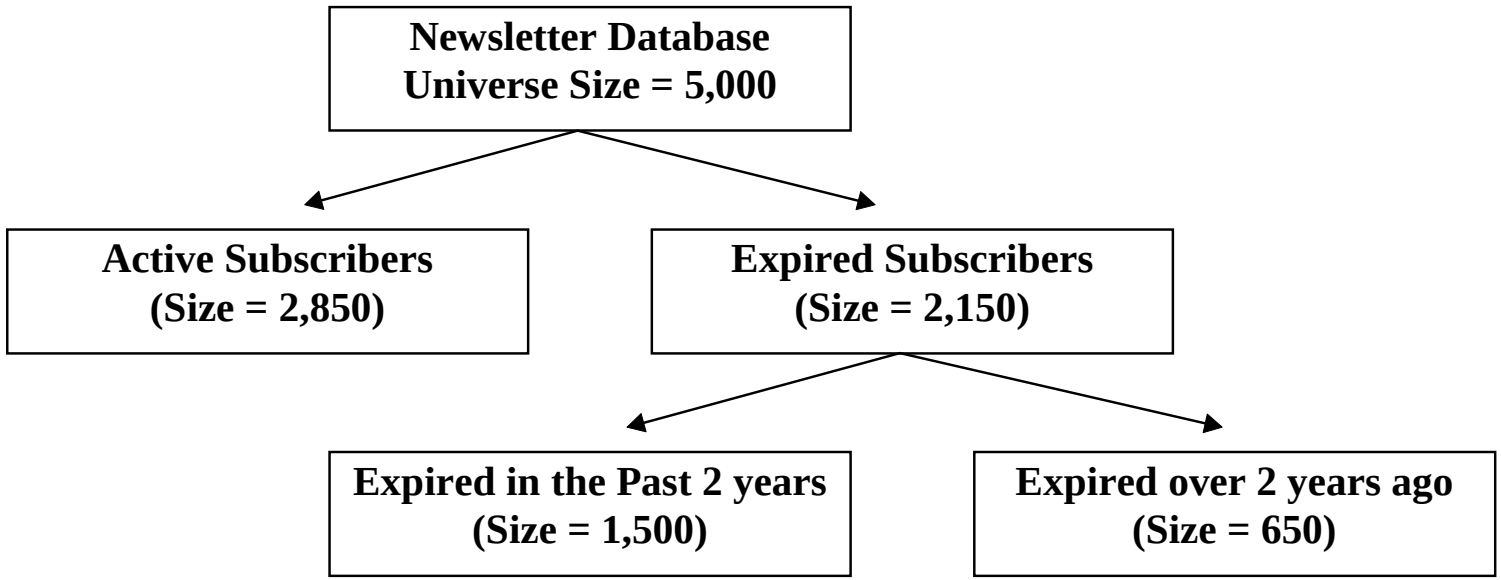
- Promotional product offers
- Life-stage marketing
- Market research

Why We Segment (Cont.)

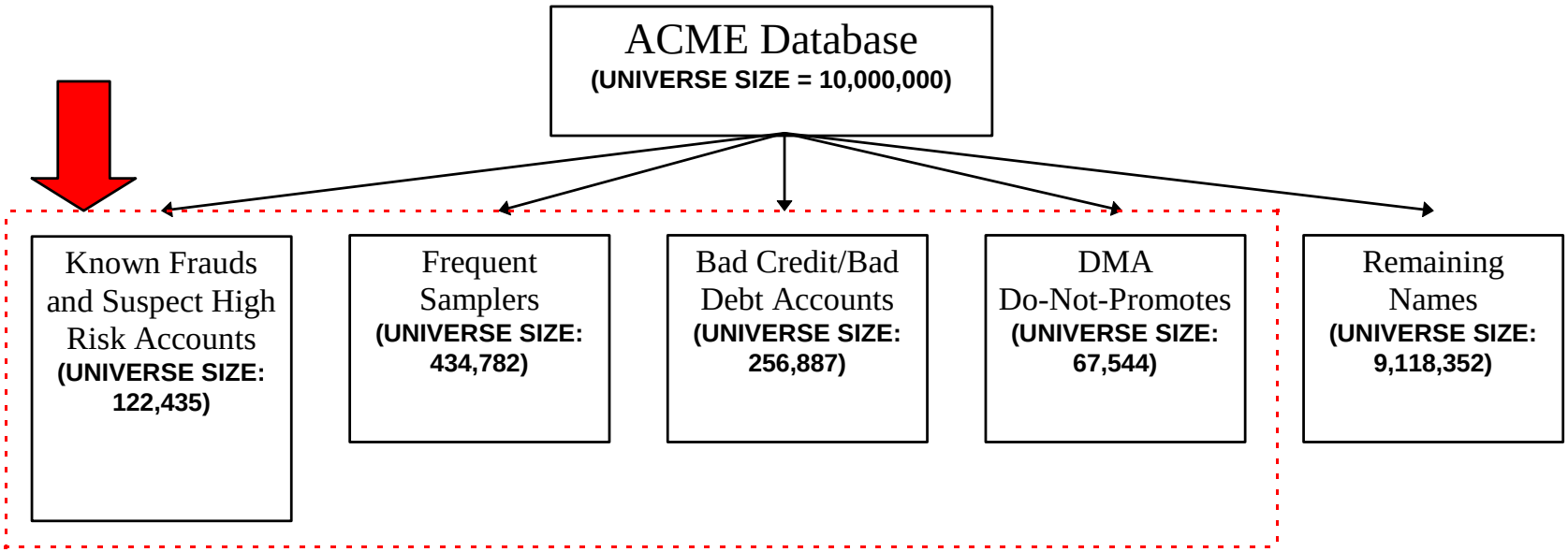
Not every segmentation scheme needs to be complex or difficult to implement. It depends on your needs as a marketer and the size of your database. Even a small home-based newsletter operation, for example, with a customer file of only 5,000 names can benefit by applying segmentation logic to their database of customers.

Why We Segment (Continued)

For example:



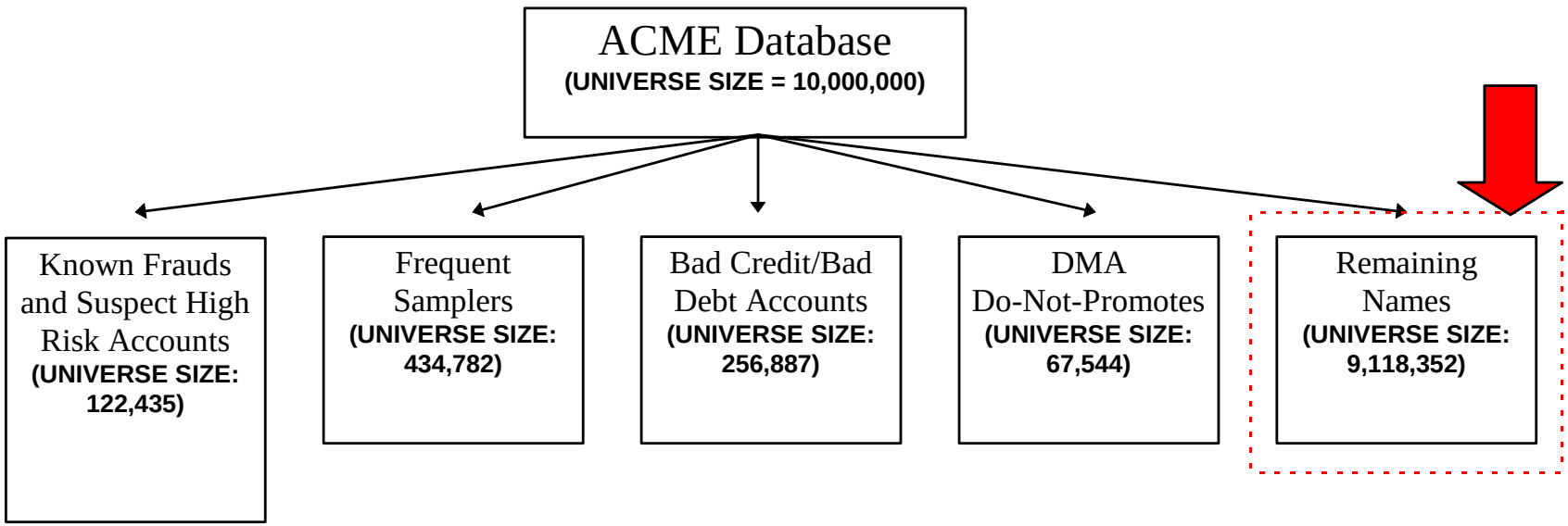
The Segmentation Scheme



Typically, corporate level eliminations are identified first and set aside for special treatment by the corporation.

The Segmentation Scheme

Once corporate level eliminations/segments have been identified, the next step is to determine how to segment the remaining names for each product line. Each product line will segment the "Remaining Names" to best meet their objectives. The most important data elements for segmenting the file for the individual product lines are recency, frequency and monetary values.



Product Line Segmentation (Continued)

Typically, a product line segmentation scheme divides the “Remaining Names” (typically called “promotable” names) into groups generically defined as the primary, secondary, tertiary, ... and finally, the conversion segment. These divisions are based on recency, frequency and monetary data related to the product line of concern:

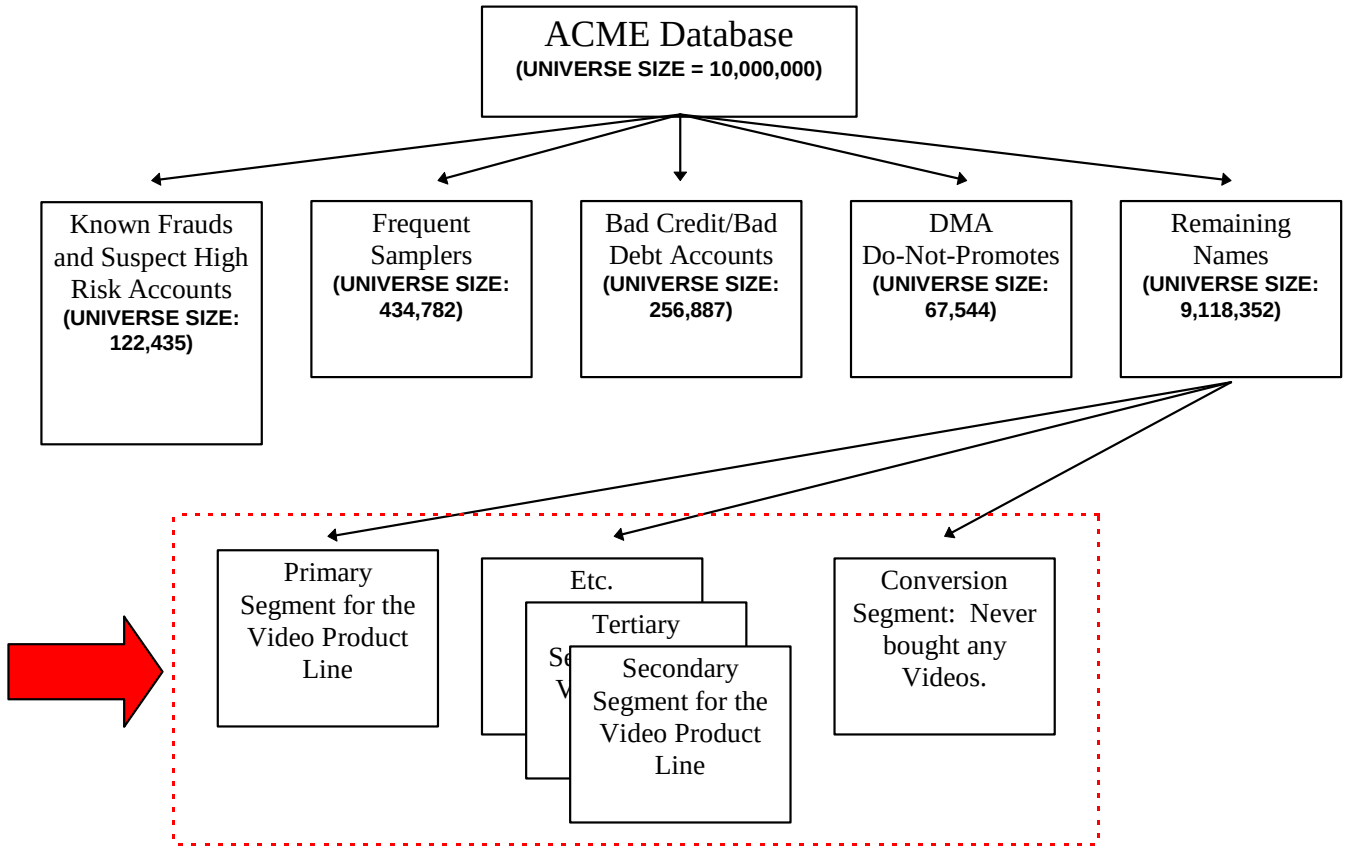
- The primary segment is considered the most profitable and generally consists of the most recently active customers for a given product line.

- The secondary segment is somewhat less profitable than the primary segment but, again, is fairly active for a given product line. Product managers employ various marketing tactics to move these names into the “primary” segment.

- The conversion segment is defined as the group of customers who have never purchased any products within the product line of interest. Product managers will employ “conversion” tactics to move these names into the “primary” segment.

Product Line Segmentation (Continued)

An example of how the ACME Direct video product line segments the "remaining names" at a high level is shown below. Unique marketing strategies can now be employed for each segment.



Segmentation for Life-Stage Marketing

Life-stage segmentation divides the file in a way that considers primarily demographic and psychographic data. This enables marketers to develop, market, or advertise more relevant products and offers on the basis of their customers' life-stages. Segmenting a customer file in this manner also allows a direct marketer to understand the future needs of their customers.

Segmentation for Life-Stage Marketing (Cont.)

Life-stages can be modeled using a multitude of internal and external enhancement data. Life-stage segments residing on a customer database might include:

- Young families with newborns
- Newly moved
- Professional 25-40 year olds, no children
- Entering the retirement years
- Children 2-5, 6-8, 9-12
- Adolescents
- College students
- New grandparents

Segmentation Techniques

There are four commonly used methods for segmenting a customer file for promotional product offers, life-stage marketing, or market research purposes:

- Univariate and cross-tabulation analysis
- - *the most common means of segmenting a customer file.*
- Formal RFM analysis*
- - *no longer considered an optimal means of segmenting the database by most analysts. But is still effective for small direct marketers.*
- CHAID analysis+
- Multivariate analysis+

* See Arthur Huges Book, “The Complete Database Marketer” for more information on this technique.

\+ See any advanced level applied statistics book for more information on these topics.

Section 3
Selecting the Test Sample

Introduction

Any new direct marketing product test or promotional test (format, creative, price) begins with taking a sample of the names residing on the database.

This sample is then sent the test promotion, product or offer. Once results of the test are final, it is this sample that will be analyzed to determine, for example, the percent responders, the payment rate, the unique characteristics separating responders from non-responders, etc. Based on this information, a marketing decision can be made.

Sample Usage

Testing is the foundation upon which direct marketing is built. With testing, a direct marketer can:

- Evaluate new product offerings
- Gauge the reaction to price changes by measuring the associated increase or decrease in response rates
- Determine the impact of a new promotional format change on response, payment or conversion rates
- Identify the target market for a new product test, for example, by reviewing the characteristics (based on internal and external data) of the responders to the product offering for any patterns that might explain why some people responded and others did not
- Gain insight about specific customer groups or segments based on customer profiles using any internal and external customer data (Section 1)

Data Residing on the Sample

The sample of customers taken from a database contains whatever information is requested by the product manager or analyst.

In most cases, a test sample will contain the same customer data that is residing and attached to each customer record on the marketing database. This includes each selected customer's promotion, purchase and payment history (internal **RFM** data) and any enhancement (external) data in addition to the customer's name and address.

Point-in-Time Data

When customer data on the test sample is saved, it must be reflective of the customer's history prior to receiving the test promotion. Think of it as a "snap shot" of each customer's record prior to sending the promotion. This "snap shot" of the customer records is also referred to as a "frozen file."

If the customer data on the sample is not reflective of the customers' "status" prior the promotion, the analysis will lead to erroneous conclusions.

Point-in-Time Data (Cont.)

There are two simple rules to remember:

- When profiling responders or determining a target market based on a test sample, customer data residing on the test sample must reflect the customer at “point-in-time” of the test promotion or, in other words, the customer’s status at the time the decision to respond or not respond to the offer is made.

- Applying the target market definition to select names for an upcoming promotion should be done with only the most current customer data available on the database. For example, after determining married customers are the best prospects for a particular promotion, assure only these customers are promoted by using the most current, up-to-date information. If the information residing on the database is too old, you may accidentally promote some that have since divorced and miss promoting those that have recently married.

Analysis & Validation Samples

Typically, before analysis of any sample is performed, the sample is spilt into two samples. One half is used to perform the analysis and the other for validating and calibrating the results of the analysis.

A sample is just that, a sample; and samples have a certain level of error associated with them. The validation sample is used to ensure the analyst does not make erroneous conclusions based on the error variance associated with the sample. For example, an analysis sample may suggest a high proportion of responders to a cookbook offer are cat owners. However, when the validation sample is considered, correlation between ordering the cookbook and cat ownership cannot be confirmed. Creating an analysis and validation sample is a very important step in the analysis process flow as illustrated in the following Figure.

Section 4
Analyzing Customer Data

Introduction

Before building a response model or segmenting your customer file, You must first become intimate with the customer data in order to exploit that data to its fullest predictive power. There are various ways in which to do this and will be discussed shortly.

In addition, you must ensure:

- the data residing on the marketing database is accurate and properly maintained.
- the data is measuring what you believe it to be measuring.

Getting to Know Your Data

Remember, the criteria for selecting names for a product promotion based on the analysis is only as good as the data residing on the database!

If the data is not properly maintained and updated, then your analysis will be misleading and result in false conclusions.

The Analysis

When preparing to build a target model or segment a customer file, there are numerous methods to assess the potential strength of various data elements or manipulate them to create stronger ones. These techniques include:

- Univariate analysis
- Cross tabulation analysis (variable interactions)
- Correlation analysis
- Manipulation techniques through the creation of:
  - Logic variables (adding elements)
  - Ratio variables (one element over the other)
  - Longitudinal variables
- CHAID analysis
- Multivariate analysis

The Analysis (Cont.)

When building target models, 50% - 75% of the analyst's time should be spent on data preparation using the techniques listed on the previous slide.

This is true even when using a "data-mining" tool. If the data is not well prepared and understood prior to "feeding" it into such a piece of software, the analysis will yield sub-optimal and potentially misleading results. Many direct marketers erroneously believe data-mining tools are "magic boxes" into which raw customer data is fed and the answer is produced. Nothing is further from the truth! Inputting "garbage data" produces "garbage results."

Univariate Tabulations

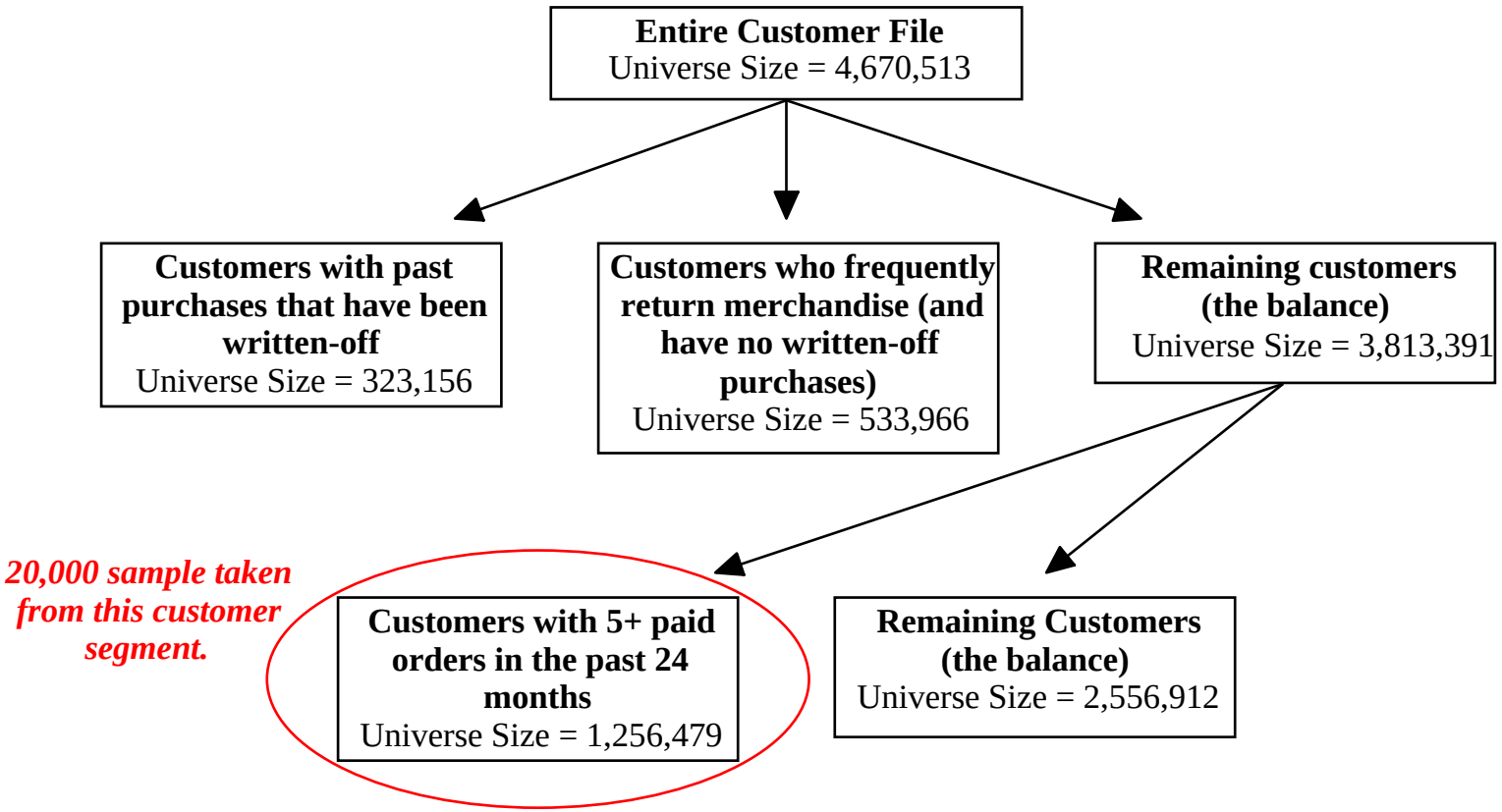
In preparing to build a target model or segment a customer file, univariate analysis is the most commonly used form of analysis. Unfortunately, it is all too often the *only* form of analysis used.

A univariate analysis involves the tabulation and viewing of a single variable or data element. In the case of a product offer test, a univariate tabulation produced on an analysis sample will display the percent responders and non-responders to the offer for each data element separately.

Univariate analysis, for example, reveals for you if customers over the age of 50 respond at a higher rate than customers under the age of 50 or if customers with 10 or more product purchases last year pay at a higher rate than customers with less than 10 product purchases.

Univariate Tabulations (Cont.)

Assume ACME Direct, a direct marketer of books, music, videos and magazines test promoted a new music CD collection titled “Pop Rock USA” (PRUSA) to 20,000 customers selected from the customer segment “5+ paid orders in the past 24 months” as shown below. All customer characteristics on the sample were frozen point-in-time of the promotion (prior section). The resulting test response rate received was 2.50%.



Univariate Tabulations (Cont.)

The sample was split 50/50 -- 10,000 names for analysis and 10,000 names for validation (prior section). The analyst was asked to examine a univariate tabulation created for the age variable to determine if any particular age group appears to distinguish responders from non-responders for this particular offer.

In the real world, an analyst will have hundreds perhaps even thousands of data elements (internal and external) to consider depending on the amount of information housed on the customer database and not just one as in this example.

Univariate Tabulations (Cont.)

An example of how the tabulation displays how “Pop Rock USA” (PRUSA) orders break out by various age categories is shown on the next slide.

The categories can vary depending on how you decide to view the data. You can have as many age categories as you like; however, keep in mind that if you break out the variable into too many categories, it may result in sparse data. Too few names in any one category will not allow you to assess the response rates with much confidence. As a result, erroneous conclusions may occur. A rule of thumb is to assess only those categories having at least 500 names falling into them. If you end up with categories having less than 500 names, consider collapsing categories. In this example, age is broken into 6 categories and each category has over 500 names.

Univariate Tabulations (Cont.)

Head of Household Age	Number	% of Sample	Number of Orders	Response Rate	Index to Total
30 and under	1,529	15.29%	67	4.38%	175
31-40	1,775	17.75%	63	3.55%	142
41-50	1,879	18.79%	46	2.45%	98
51-60	2,054	20.54%	29	1.41%	56
61 and over	1,785	17.85%	18	1.01%	40
No age info available	978	9.78%	27	2.76%	110
Total	10,000	100.00%	250	2.50%	100

The column titled “Number” represents the number of names in the sample of 10,000 falling into each age category. For example, in this sample 1,529 people out of 10,000 are 30 years of age and under.

The “% of Sample” represents the percent of the corresponding “Number” as it relates to the total sample size. In this case, 1,529 represents 15.29% of the total.

Univariate Tabulations (Cont.)

Head of Household Age	Number	% of Sample	Number of Orders	Response Rate	Index to Total
30 and under	1,529	15.29%	67	4.38%	175
31-40	1,775	17.75%	63	3.55%	142
41-50	1,879	18.79%	46	2.45%	98
51-60	2,054	20.54%	29	1.41%	56
61 and over	1,785	17.85%	18	1.01%	40
No age info available	978	9.78%	27	2.76%	110
Total	10,000	100.00%	250	2.50%	100

The “Number of Orders” column represents how many of the total 250 orders for this particular product offer fell into each category. In this example, 67 of the 250 total orders for this product were placed by individuals 30 years of age and under.

The response rate to this product offer is 4.38% (67/1,529) for those customers 30 years of age and under.

Univariate Tabulations (Cont.)

Head of Household Age	Number	% of Sample	Number of Orders	Response Rate	Index to Total
30 and under	1,529	15.29%	67	4.38%	175
31-40	1,775	17.75%	63	3.55%	142
41-50	1,879	18.79%	46	2.45%	98
51-60	2,054	20.54%	29	1.41%	56
61 and over	1,785	17.85%	18	1.01%	40
No age info available	978	9.78%	27	2.76%	110
Total	10,000	100.00%	250	2.50%	100

The index value associated for those customers 30 years of age and under is calculated by taking the response rate of this group and dividing by the response rate for the entire sample and multiplying by 100. An index value of 175 for this group tells you that the response rate for this age group is 75% higher than the response rate for all names. In other words, by promoting only those that are “30 and under,” you will obtain a response rate 75% higher than if you promoted the entire sample (4.38% vs. 2.50%). This 75% figure is called the “gain in response.”

The index value of 40 associated with the “61 and over” group implies that their response rate is 40% lower than the response rate for all names (1.01% vs. 2.50%).

Univariate Tabulations (Cont.)

Head of Household Age	Number	% of Sample	Number of Orders	Response Rate	Index to Total
30 and under	1,529	15.29%	67	4.38%	175
31-40	1,775	17.75%	63	3.55%	142
41-50	1,879	18.79%	46	2.45%	98
51-60	2,054	20.54%	29	1.41%	56
61 and over	1,785	17.85%	18	1.01%	40
No age info available	978	9.78%	27	2.76%	110
Total	10,000	100.00%	250	2.50%	100

Assuming the break-even for this particular product offering is a response rate of at least 3.00%, which names could the product manger profitably promote using age information alone?

Univariate Tabulations (Cont.)

NOTE: The “selection” criteria depends upon the objective of the promotion. For name acquisition promotions, a direct marketer may actually be willing to *lose* money on the initial promotion in order to generate customers. Publishers, having a subscriber rate base to meet, will promote as many names as needed to meet their rate base; the goal being to promote the best names. Publishers may actually *lose* up to $10 or more per subscriber in order to meet its rate base, the loss offset by advertising revenue.

Cross-Tabulations

Cross-tabulations are a means of viewing two or more univariate data elements in combination. “Cross-tabbing” highlights interrelationships among variables. It can take a relatively weak data element, in terms of it predictive strength, and change it into a power predictor.

For example, a cross tabulation of age information with gender information may reveal that those under the age of 40 do not respond at our required 3.00% level if they are female but only male.

Correlation Analysis

Correlation analysis is another means of assessing which continuous* data elements might be good predictors of the customer behavior of interest. The correlation coefficient is a measure of the relationship between two variables.

Correlation analysis can answer questions such as:

- Is the data element “total number of books purchased in the past 12 months” highly correlated with a customer’s likelihood of ordering a new video series?

- Is income level positively correlated with the amount a customer spends on a catalogue promotion? In other words, do customers with higher incomes spend more on catalogue orders?

* Or quantitative discrete data with many possible values of an ordinal nature.

Logic Variables

When you have several data elements all measuring the same thing about a customer, consider the creation of a logic variable. For example, if you have several questionnaire data elements on your database all measuring ones interest in a specific area, consider combining them and creating a logic variable. The resulting variable will be stronger than each component variable on their own and help you better assess a customer's true interest in that area.

For example, consider the following three cooking questions on your database: Do you like cooking? Do you buy cookbooks? Do you subscribe to cooking magazines?

Customer Smith replied yes to all three questions regarding cooking while customer Jones only replied yes to one of them. Which do you think is most interested in cooking?

Ratio Variables

Ratio variables are the result of dividing one data element by another. The data elements comprising the ratio variable must be continuous in nature.

For example, instead of considering each of the data elements “Total Promotions Sent” and “Total Orders Placed” separately, combine them into the ratio “Orders/promotions.”

It will yield an average order rate over time for each customer and be quite predictive of future customer actions.

Ratio Variables

Below are examples of ratio variables created from data on ACME Direct's customer file.

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, videos, gifts, etc.)
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

Ratio Variables

Below is a partial printout of three customers from the 10,000-name analysis sample tested for “Pop Rock USA.” All data is representative of the point-in-time of the promotion. Based on “Total Orders” information alone, which of the three customers is most likely to respond to this offer?

Customer Name	Customer Address	Total Promotions	Total Orders
T. Bluestone	555 Maple		10
R. Stewart	56 South Main		7
J. Jackson	111 Rocky Rd.		2

Ratio Variables

Based on “Total Promotions” information alone shown below who is more likely to respond to this offer? (Assume you were forced to make a selection based on this information alone.)

Customer Name	Customer Address	Total Promotions	Total Orders
T. Bluestone	555 Maple	84	
R. Stewart	56 South Main	55	
J. Jackson	111 Rocky Rd.	12	

Ratio Variables

Consider the ratio of “Total Orders” to “Total Promotions” for each customer and determine who is more likely to respond to this offer. The creation of the ratio variable is shown below.

Customer Name	Customer Address	Total Promotions	Total Orders	Ratio of Orders to Promotions
T. Bluestone	555 Maple	84	10	10/84 = 11.90%
R. Stewart	56 South Main	55	7	7/55 = 12.73%
J. Jackson	111 Rocky Rd.	12	2	2/12 = 16.67%

Longitudinal Variables

Longitudinal variables allow a direct marketer to view a particular data element for each customer across time. Conceptually, they are similar to time-series variables and can be quite difficult to implement.

Longitudinal variables are based on the premise that the best predictor of a customer's future response to a promotion is a review of his most recent past responses and reactions to promotions.

For example, longitudinal data will allow you to consider information such as each customers last three actions to promotions sent

Longitudinal Variables

The figure below displays some examples of longitudinal/time-series variables.

LONGITUDINAL VARIABLE	*MEASUREMENT OF*
Customer's response (order, pay, silent, etc.) to their last 3 promotions sent	Estimate of customer's action on next promotion sent
Customer's action (pay, return, bad-debt) to their last 3 orders placed	Estimate of customer's performance on next order placed
Customer's last 3 product affinity shipments for the music club (rock, pop, country, etc.)	Estimate of customer's most current music interests

Section 5

Regression Modeling

Introduction

In an attempt to determine the most profitable names to promote for an upcoming campaign, nothing will be more effective than the development and deployment of a response model.

Such a model will consider many customer attributes in combination to assist you in determining who to promote and not promote for you campaigns.

Customer Actions Typically Modeled

Customer actions typically modeled in the direct marketing industry include:

- An order
- A bad debt/account write-off (for those direct marketers extending free-trial offers)
- A return/cancel to the initial offer
- A paid order (a combined order and payment model)
- A renewal the next year out (publishing industry)
- Subsequent shipments (for clubs)
- Catalogue expenditures

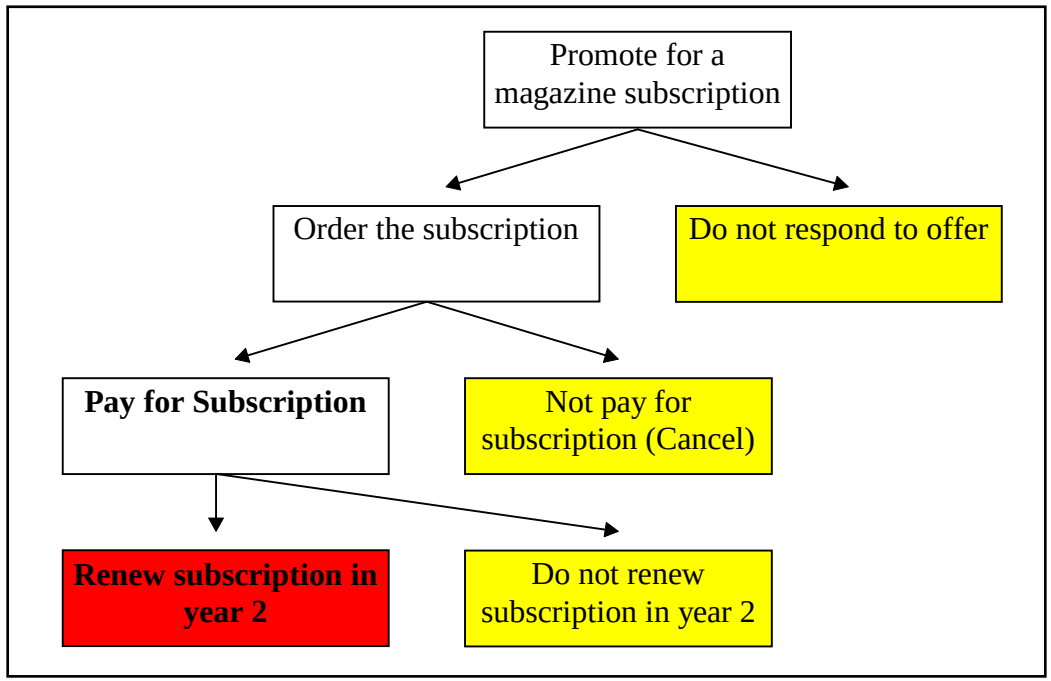
Customer Actions Typically Modeled

For example, in determining who to promote for an upcoming magazine subscription offer, a direct marketer may decide to build several models predicting:

- How likely someone is to order the subscription
- How likely someone is to not cancel the subscription given they have ordered
- How likely someone is to renew the subscription the next year.

Customer Actions Typically Modeled

A direct marketer will promote those names most likely to fall within the red cell of the figure above and not promote those names most likely to fall within the yellow cells. Multiple regression models will help you determine how likely a customer is to fall within each of these cells.



The Multiple Regression Model

Assuming ample time was spent massaging and selecting powerful predictor variables, the modeling process itself is quite simple. The form of a multiple regression model is:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots$$

Where:

- y is the response variable, what you are predicting (e.g., an order, a cancel, etc.)
- $x_1, x_2, x_3, \ldots$ are the multiple predictor variables (e.g., age, income, RFM data elements, etc.)
- a is a constant numerical value
- $b_1, b_2, b_3, \ldots$ are the numerical coefficients (weights) associated with each of the predictor variables

Model Interpretation

We will now examine what an actual multiple regression model looks like. Consider the following example: ACME Direct test promoted 20,000 names selected at random from the database a rock music package last year.

- The sample was saved including all customer attributes point-in-time of the promotion.
- The ACME Direct analyst was asked by the product manager to build a multiple regression model that would help identify those most likely to order this country music package.
- The ACME Direct analyst split the saved sample: 10,000 for analysis and 10,000 for validation.
- Using the analysis sample, he created the *binary dependent* response variable "Order" which took on the two values: 0 if the customer did not order the music package and 1 if the customer ordered the music package.
- He then examined the data and identified customer attributes that appeared to be strong predictors of ordering this particular product using the techniques briefly discussed in the prior section including univariate and cross tabulations.

The final multiple regression model is shown on the next slide.

Model Interpretation (Continued)

Variable	Definition	Coefficient/Weight
Constant	A constant value	0.151767
X1	=1 if activity of any kind in the past 6 months = 0 otherwise	0.023618
X2	= number of music products paid in the past 24 months	0.060634
X3	= 1 if indicated they like rock music on questionnaire = 0 otherwise	0.008761
X4	= ratio of rock music paid to rock music promotions sent	0.525923
X5	= 1 if bought one or more of the following four rock music titles: Rock and Roll Party, Soul of Rock and Roll, Early Rock Legends, Easy Listening Rock = 0 otherwise	0.086853

Model Interpretation (Continued)

Variable	Definition	Coefficient/Weight
Constant	A constant value	0.151767
X1	=1 if activity of any kind in the past 6 months = 0 otherwise	0.023618
X2	= number of music products paid in the past 24 months	0.060634

In general, think of a multiple regression model as a set of specific criteria used to create a "score" for each customer that will indicate their relative likelihood of ordering (or paying, canceling, renewing, etc.) the product of concern. Some characteristics/attributes will add to a customer's score while others lower the score. The way in which each variable affects a customers overall value when scored is shown on the next slide. The higher the score (the closer to a value of 1) the more likely the customer will order.

Model Interpretation (Continued)

Variable	Definition	Coefficient/ Weight	Application Scoring Rules
Constant	A constant value	0.151767	Every customer "scored" on the regression model begins with a score equal to the constant value of 0.151767
X1	= 1 if activity of any kind in the past 6 months = 0 otherwise	0.023618	If a customer has exhibited any activity in the past 6 months, their score is increased by 0.023618; otherwise, their score is unaffected.
X2	= number of music products paid in the past 24 months	0.060634	Each customer's score is increased by the number of music products paid in the past 24 months when multiplied by 0.060634
X3	= 1 if indicated they like rock music on questionnaire = 0 otherwise	0.008761	If a customer said they like rock music on the questionnaire, their score is increased by 0.008761; otherwise their score is unaffected.
X4	= ratio of rock music paid to rock music promotions sent	0.525923	Each customer's score is increased by the value of the ratio of rock music paid orders to rock music promotions sent when multiplied by 0.525923.
X5	= 1 if bought one or more of the following four rock music titles: Rock and Roll Party, Soul of Rock and Roll, Early Rock Legends, Easy Listening Rock = 0 otherwise	0.086853	If a customer bought 1 or more of the four rock music titles listed, their score is increased by 0.086853; otherwise their score is unaffected.

Model Interpretation (Continued)

Below are two customers from the ACME Direct database along with relevant data. Assume today's date is 3/1/99. Which customer, Smith or Jones, is more likely to order the country music package based on the multiple regression model?

Customer	Last Activity Date	Music Paid in the Past 24 Months	Rock Music Paid Ever	Rock Music Promotions Ever	Question - Like Rock Music?	Rock Music Purchase History			
						RRP	SRR	ERL	ELR
Smith	8/1/98	4	0	7	No	Yes	No	Yes	No
Jones	11/7/98	2	1	6	Yes	No	No	No	No

Model Interpretation (Continued)

To determine who is more likely to order this product, we score both Jones and Smith on the regression model as follows:

Variable	Smith Status	Smith Score		Jones Status	Jones Score
Constant	yes	.151767		yes	.151767
Act w/in 6 months	no	0		yes	.023616
Music pd w/in 24 months	4	4x(.060634) = .242536		2	2x(.060634) = .121268
Question - like country music	no	0		yes	.008761
Ratio rock paid to promotions sent	0/7=0	0		1/6 = .16667	.16667x(.525923 = .087671
Paid 1+ rock titles	yes	.086853		no	0
TOTAL SCORE		**0.481156**			**0.393083**

Regression Diagnostics

When building a response model there are many assumptions of the model that must not be violated. If violated, the model will be unstable and non-predictive. The model will not hold up in roll-out.

For more information on these assumptions view any intermediate to advanced level applied statistics book such as “Applied Multivariate Statistical Analysis” by Johnson and Wichern (Prentice Hall Publishers).

Assumption of the Model

The single most important and most easily understood assumption regarding the build of a regression model is the *assumption of independence*.

The *assumption of independence* states that:

Each of the predictor variables used in the final regression model must be independent of one another.

In other words, any two predictor variables in the final model cannot be correlated with one another. Each predictor must be measuring something unique about the customer behavior of interest.

A violation of this assumption is called *multicollinearity*. If a model possesses multicollinearity, it will be unstable and unpredictable when applied in a roll-out situation.

Assumption of the Model (Continued)

For example, would it be wise to consider
using both “household income” and “home value” as separate predictors
in a multiple regression model?

How could you safely consider using both
measures of wealth in a multiple regression model?

Section 6

The Gains Chart

Introduction

Up to this point we have focused on the customer data and modeling customer behavior using that data.

Now we will learn how to validate the model and prepare to use it to help us determine who to promote and not promote for an upcoming direct mail campaign.

The Steps of the Process

Once the model is built the first thing to be done is to validate the model to ensure it will hold up in a roll-out situation. This will de done with the use of a gains chart.

Once the model is validated via the gains chart the gains chart will be used to determine who to promote and not promote for a particular offer.

The Gains Chart

How we measure the strength of one binary response model over another in the field of direct marketing is by the creation of the gains chart not the R^2 value.

The Gains Chart

The steps taken to develop a gains chart are:

STEP 1: The analysis sample is scored on the final regression model.

STEP 2: The scored sample is ranked from highest to lowest in terms of the regression scores.

STEP 3: The ranked sample is then cut into "buckets." For example, the top 10%, the next highest scoring 10%, and so on. How many buckets one uses depends on how large the sample is and how many orders are in the sample. Typically one uses anywhere from 10 to 50 buckets.

STEP 4: Once the buckets are determined, calculate the actual response rate for each bucket based on how many of the orders from the sample fell into each bucket.

The Gains Chart

STEP 5: The gain for each bucket is calculated by determining the index associated with that bucket when compared to the response rate of the entire sample and subtracting a value of one then multiplying it by 100. For example, if the top 10% (or decile) of the names when ranked on the regression score had a response rate of 15.26% and the response rate for the entire sample was 4.67%, the gain in response associated with this bucket of names would be determined as follows:

Gain = ((15.26%/4.67%) - 1) x 100

= ((3.27) - 1) x 100

3.27 is called the index to total for this bucket of names.

= (2.27)x100

= 227 ***This value tells you that this particular bucket of names has a response rate 227% higher than the entire sample.***

Example

ACME Direct test promoted an offer for a new magazine called "Internet Communications." The test was comprised of 20,019 names promoted from the entire database of known computer owners less corporate level eliminations (frauds, DMA non-promotes, etc.).

The overall response rate for the test was 6.74%.

Based on the success of the test, ACME has decided to go forward and offer this new magazine to those customers on it's database who are most likely to respond. ACME Direct will determine those on it's database that are most likely to respond to this offer by building a response model.

Example

The ACME analyst first divides the sample of 20,019 names into two parts - one for analysis and the other for validation. The analyst then builds a regression model predicting who is most likely to order this magazine subscription based on the analysis portion of the sample.

Once the analyst is satisfied with the model, he will score the analysis portion of the sample the model was built on and develops a 10 bucket gains chart where each bucket represents 10% of the analysis sample.

Example (Continued)

Incremental figures are the figures associated with only that bucket of names.

Cumulative figures are associated with that bucket of names and above.

The sample was cut into 10 - 10% buckets

ANALYSIS GAINS CHART FOR "INTERNET COMMUNICATIONS"

		Incremental				Cumulative			
Bckt	Score	Pct.	Num.	Resp. Rate	Gain	Pct.	Num.	Resp. Rate	Gain
1	GE .3543	10.09%	1,010	12.67%	88	10.09%	1,010	12.67%	88
2	.3017 - .3542	9.91%	992	10.87%	61	20.00%	2,002	11.78%	75
3	.2598 - .3016	10.00%	1,001	9.45%	40	30.00%	3,003	11.00%	63
4	.2291 - .2597	10.20%	1,021	8.55%	27	40.20%	4,024	10.38%	54
5	.2004 - .2290	9.80%	981	7.05%	5	50.00%	5,005	9.73%	44
6	.1832 - .2003	10.00%	1,001	5.93%	-12	60.00%	6,005	9.09%	35
7	.1644 - .1831	10.05%	1,006	4.76%	-29	70.05%	7,011	8.47%	26
8	.1599 - .1643	9.95%	996	3.54%	-47	80.00%	8,007	7.86%	17
9	.1274 - .1598	10.00%	1,001	2.89%	-57	90.00%	9,008	7.31%	8
10	LE 0.1273	10.00%	1,001	1.59%	-76	100.00%	10,009	6.74%	0
ALL		100%	10,009	6.74%	100				

Obviously there were some tie scores at the bucket cuts

The response rates calculated for each bucket of names.

Nice and smooth monotonically decreasing incremental gains!

Example (Continued)

ANALYSIS GAINS CHART FOR "INTERNET COMMUNICATIONS"

		Incremental				Cumulative			
Bckt	Score	Pct.	Num.	Resp. Rate	Gain	Pct.	Num.	Resp. Rate	Gain
1	GE .3543	10.09%	1,010	12.67%	88	10.09%	1,010	12.67%	88
2	.3017 - .3542	9.91%	992	10.87%	61	20.00%	2,002	11.78%	75
3	.2598 - .3016	10.00%	1,001	9.45%	40	30.00%	3,003	11.00%	63
4	.2291 - .2597	10.20%	1,021	8.55%	27	40.20%	4,024	10.38%	54
5	.2004 - .2290	9.80%	981	7.05%	5	50.00%	5,005	9.73%	44
6	.1832 - .2003	10.00%	1,001	5.93%	-12	60.00%	6,005	9.09%	35
7	.1644 - .1831	10.05%	1,006	4.76%	-29	70.05%	7,011	8.47%	26
8	.1599 - .1643	9.95%	996	3.54%	-47	80.00%	8,007	7.86%	17
9	.1274 - .1598	10.00%	1,001	2.89%	-57	90.00%	9,008	7.31%	8
10	LE 0.1273	10.00%	1,001	1.59%	-76	100.00%	10,009	6.74%	0
ALL		100%	10,009	6.74%	100				

Suppose the goal of this mailing is to achieve at least 10% profit after overhead. If you know that in order to obtain this profit level, you need at least an 8% response rate, what buckets would you mail?

Example (Continued)

ANALYSIS GAINS CHART FOR "INTERNET COMMUNICATIONS"

		Incremental				Cumulative			
Bckt	Score	Pct.	Num.	Resp. Rate	Gain	Pct.	Num.	Resp. Rate	Gain
1	GE .3543	10.09%	1,010	12.67%	88	10.09%	1,010	12.67%	88
2	.3017 - .3542	9.91%	992	10.87%	61	20.00%	2,002	11.78%	75
3	.2598 - .3016			9.45%	40	30.00%	3,003	11.00%	63
4	.2291 -			8.55%	27	40.20%	4,024	10.38%	54
5	.2004 - .		981	7.05%	5	50.00%	5,005	9.73%	
6	.1832 - .2003	10.00%	1,001	5.93%	-12	60.00%	6,005		
7	.1644 - .1831	10.05%	1,006	4.76%	-29				
8	.1599 - .1643	9.95%	996	3.54%	-47				
9	.1274 - .1598	10.00%	1,001	2.89%	-57			7.31%	8
10	LE 0.1273	10.00%	1,001	1.59%	-76		10,009	6.74%	0
ALL		100%	10,009	6.74%	100				

MAIL THIS BUCKET AND ABOVE

YOUR FORECASTED MAIL QUANTITY PERCENT, RESPONSE RATE AND OVERALL GAIN

You use the incremental column to determine who to mail (your cut-off level). You use the cumulative column to determine your forecasted mail quantity and response rate at that cut off level.

The Validation Gains Chart

Remember, as was told to you back in Section 3, we should not base our marketing decisions on the analysis sample but rather the validation (or hold-out) sample.

The main reason is because the model built on the analysis sample will tend to over-predict. The validation sample helps you overcome this problem and will yield results more closely resembling what you will obtain in roll-out.

To develop a validation gains chart you apply the following steps.

The Validation Gains Chart (Continued)

The steps taken to develop a validation gains chart are:

STEP 1: The validation sample is scored on the final regression model.

STEP 2: The scored sample is then cut into the buckets as defined by the analysis sample.

STEP 3: Once the buckets are determined, you then calculate the response rate in each.

STEP 4: The gains are calculated for each bucket.

Example (Continued)

Going back to our previous example, the analyst scores the validation sample using the "Internet Communications" response model that was built on the analysis sample and builds a validation gains chart.

Following are the gains charts for both the analysis and validation samples.

Example (Continued)

The validation sample is cut on the same scores determined on the analysis sample.

ANALYSIS AND VALIDATION CUMULATIVE GAINS CHARTS FOR "INTERNET COMMUNICATIONS"

		ANALYSIS SAMPLE			VALIDATION SAMPLE		
Bucket	Score	Percent	Resp. Rate	Gain	Percent	Resp. Rate	Gain
1	GE .3543	10.09%	12.67%	88	10.00%	12.04%	79
2	.3017 - .3542	20.00%	11.78%	75	20.00%	11.33%	68
3	.2598 - .3016	30.00%	11.00%	63	30.15%	10.68%	58
4	.2291 - .2597	40.20%	10.38%	54	40.00%	10.17%	51
5	.2004 - .2290	50.00%	9.73%	44	50.00%	9.59%	42
6	.1832 - .2003	60.00%	9.09%	35	60.02%	8.99%	33
7	.1644 - .1831	70.05%	8.47%	26	70.00%	8.41%	25
8	.1599 - .1643	80.00%	7.86%	17	80.00%	7.84%	16
9	.1274 - .1598	90.00%	7.31%	8	90.34%	7.30%	8
10	LE 0.1273	100.00%	6.74%	0	100.00%	6.74%	0
ALL							

Once the validation sample is cut, the percents, response rates and gains for each bucket are calculated.

Notice the percents falling into each bucket (as defined by the analysis sample) are not exactly the same for the validation sample.

Example (Continued)

ANALYSIS AND VALIDATION CUMULATIVE GAINS CHARTS FOR "INTERNET COMMUNICATIONS"

		ANALYSIS SAMPLE			VALIDATION SAMPLE			
Bucket	Score	Percent	Resp. Rate	Gain	Percent	Resp. Rate	Gain	
1	GE .3543	10.09%	12.67%	88	10.00%	12.04%	79	**-10.6%**
2	.3017 - .3542	20.00%	11.78%	75	20.00%	11.33%	68	**-8.9%**
3	.2598 - .3016	30.00%	11.00%	63	30.15%	10.68%	58	**-7.5%**
4	.2291 - .2597	40.20%	10.38%	54	40.00%	10.17%	51	**-5.8%**
5	.2004 - .2290	50.00%	9.73%	44	50.00%	9.59%	42	**-4.7%**
6	.1832 - .2003	60.00%	9.09%	35	60.02%	8.99%	33	**-4.3%**
7	.1644 - .1831	70.05%	8.47%	26	70.00%	8.41%	25	**-3.5%**
8	.1599 - .1643	80.00%	7.86%	17	80.00%	7.84%	16	**-1.8%**
9	.1274 - .1598	90.00%	7.31%	8	90.34%	7.30%	8	**-1.7%**
10	LE 0.1273	100.00%	6.74%	0	100.00%	6.74%	0	**N/A**
ALL								

Notice how our cumulative gains have decreased from analysis to validation. The lose in gains for the top bucket is -10.6%. Typically this loss in gains lessens the deeper you go in the gains chart (as shown here).

Gains Fall-off

Typically, you should not see more than a 10% fall off in gains in the top buckets from analysis to validation. If you see significantly more fall off than this, it may indicate a problem with the model built:

- **Marginally significant predictors based on the p-values***
- **Multicollinearity***

**** For more information on these topics view any intermediate to advanced level applied statistics book.***

Expected Profit Gains Chart

The gains chart just examined was produced on each customers score representing their overall likelihood of ordering. Another type of gains chart that allows more precision in selecting names is one produced on each customers expected value to the corporation if promoted.

To create such a gains chart you will not only consider the likelihood that each customer will order but also the likelihood that each customer will pay in addition to the marketing cost and profit values associated with the promotion.

Expected Profit Gains Chart

For example, you are promoting a magazine subscription offer for which the cancel rate is high. In determining who to promote, you build both a response model and a payment model to help you select not only customers most likely to order but customers most likely to order and pay.

This is where the calculation of expected profit comes into play.

Expected Profit Gains Chart

To determine the expected profit or loss for a particular business scenario you multiply the probabilities associated with each possible outcome by their respective net costs or profit values, and then sum.

The Expected Monetary Value calculation (EMV) is written as:

$$EMV = P(O_1)M_1 + P(O_2)M_2 + P(O_3)M_3 + \ldots + P(O_n)M_n$$

Where:

$P(O_1)$, $P(O_2)$, $P(O_3)$, … $P(O_n)$ equal the probabilities associated with each of the *n* possible outcomes of the business scenario and the sum of these probabilities must equal 1.

M_1, M_2, M_3, … M_n, equals the net monetary values (costs or profit values) associated with each of the *n* possible outcomes of the business scenario.

Expected Profit Gains Chart

The easiest way to understand EMV is to review a lottery example. Assume you are considering the purchase of a lottery ticket where the probability of winning the $1 million jackpot is 1 in 10,000,000. If the lottery ticket costs $1 to purchase, what is the expected monetary value or expected profit/loss for this scenario (purchasing of a lottery ticket)?

Expected Profit Gains Chart

First, list the various outcomes (O_1, O_2, O_3, … On) of this scenario. In this case there are only 2 (n = 2):

Possible Outcomes
O_1 = You win the lottery
O_2 = You don't win

Next, determine the probabilities $P(O_1)$ and $P(O_2)$ associated with each of the two outcomes:

Possible Outcomes	**Probability**
O_1 = You win the lottery	$P(O_1)$ = 1/10,000,000 = .0000001
O_2 = You don't win	$P(O_2)$ = 9,999,999/10,000,000 = .9999999

Expected Profit Gains Chart

Now, determine the associated net monetary values M_1 and M_2 for each of the two outcomes:

Possible Outcomes	**Probability**	**Monetary Value**
O_1 = You win the lottery	$P(O_1) = .0000001$	$M_1 =$ \$999,999
O_2 = You don't win	$P(O_2) = .9999999$	$M_2 =$ -\$1

Note: The net monetary value associated with winning the lottery is calculated by taking the winnings of \$1,000,000 and subtracting the \$1 cost of the ticket. The net value is \$999,999.

Expected Profit Gains Chart

Finally, calculate the EMV as:

$$\begin{aligned} EMV &= P(O_1)M_1 + P(O_2)M_2 \\ &= (.0000001)(\$999{,}999) + (.9999999)(-\$1) \\ &= \$0.0999999 - \$0.9999999 \\ &= -\$0.90 \end{aligned}$$

In other words, if you play this particular lottery game, your expected payout is a loss of 90 cents. Keep in mind that this is an average value meaning that if you play this particular game over and over again, your average payout per play is a loss of 90 cents.

To learn more about these topic plus many others, you can purchase Perry's book

"Optimal Database Marketing"

by Sage Publications at Amazon.com.