The Charles Book Club

This case has been prepared by Ms. Vinni Bhandari, a data mining consultant and Dr. Nitin Patel, a visiting professor of Operations Research at the MIT Sloan School of Management. The case has been derived from 'BBB - The Bookbinders Club, 'a Case Study in Database Marketing, prepared by Nissan Levin and Jacob Zahavi, Tel Aviv University for the Direct Marketing Educational Foundation, Inc. (March 1995).

THE BOOK INDUSTRY

Approximately 50,000 new titles, including new editions, are published each year in the US, giving rise to a $25 billion industry in 2001.[1] In terms of percentage of sales, this industry may be segmented as follows:

16%	Textbooks
16%	Trade books sold in bookstores
21%	Technical, scientific and professional books
10%	Book clubs and other mail-order books
17%	Mass-market paperbound books
20%	All other books

Book retailing in the US in the 1970's was characterized by the growth of bookstore chains located in shopping malls. The 1980's saw increased purchases in bookstores stimulated through the widespread practice of discounting. By the 1990's, the superstore concept of book retailing gained acceptance and contributed to double-digit growth of the book industry. Conveniently situated near large shopping centers, superstores maintain large inventories of 30,000 to 80,000 titles, and employ well-informed sales personnel. Superstores applied intense competitive pressure on book clubs and mail-order firms as well as traditional book retailers. In response to these pressures, book clubs sought out alternative business models that were more responsive to their customers' individual preferences.

Historically, book clubs offered their readers different types of membership programs. Two common membership programs are "continuity" and "negative option" programs that were extended contractual relationships between the club and its members. Under a continuity program, a reader would sign up by accepting an offer of several books for just a few dollars (plus shipping and handling) and an agreement to receive a shipment of one or two books each month thereafter at more standard pricing. The continuity program was most common in the children's books market, where parents are willing to delegate the rights to the book club to make a selection, and much of the club's prestige depends on the quality of its selections. In a negative option program, readers get to select which and how many additional books they would like to receive. However, the club's selection of the month will be automatically delivered to them unless they specifically mark "no" by a deadline date on their order form. Negative option programs sometimes result in customer dissatisfaction and always give rise to significant mailing and processing costs.

In an attempt to combat these trends, some book clubs have begun to offer books on a “positive option” basis, but only to specific segments of their customer base that are likely to be receptive to specific offers. Rather than expanding the volume and coverage of mailings, some book clubs are beginning to use database-marketing techniques to more accurately target customers. Information contained in their databases is used to identify who is most likely to be interested in a specific offer. This information enables clubs to carefully design special programs tailored to meet their customer segments' varying needs.

DATABASE MARKETING AT CHARLES

The club

The Charles Book Club ("CBC") was established in December of 1986, on the premise that a book club could differentiate itself through a deep understanding of its customer base and by delivering uniquely tailored offerings. CBC focused on selling specialty books by direct marketing through a variety of channels, including media advertising (TV, magazines, newspapers) and mailing. CBC is strictly a distributor and does not publish any of the books that it sells. In line with its commitment to understanding its customer base, CBC built and maintained a detailed database about its club members. Upon enrollment, readers were required to fill out an insert and mail it to CBC. Through this process, CBC has created an active database of 500,000 readers. CBC acquired most of these customers through advertising in specialty magazines.

The problem

CBC sent mailings to its club members each month containing its latest offering. On the surface, CBC looked like they were very successful, mailing volume was increasing, book selection was diversifying and growing, their customer database was increasing; however, their bottom line profits were falling. The decreasing profits led CBC to revisit their original plan of using database marketing to improve its mailing yields and to stay profitable.

A possible solution

They embraced the idea that deriving intelligence from their data would allow them to know their customer better and enable multiple targeted campaigns where each target audience would receive appropriate mailings. CBC's management decided to focus its efforts on the most profitable customers and prospects, and to design targeted marketing strategies to best reach them. The two processes they had in place were:

1. Customer acquisition:
   - New members would be acquired by advertising in specialty magazines, newspapers and TV.
   - Direct mailing and telemarketing would contact existing club members.
   - Every new book would be offered to the club members before general advertising.

2. Data collection:
   - All customer responses would be recorded and maintained in the database.
   - Any information not being collected that is critical would be requested from the customer.

To derive intelligence from these processes they decided to use a two-step approach for each new title:

a. Conduct a market test, involving a random sample of 4,000 customers from the database to enable analysis of customer responses. The analysis would create and calibrate response models for the current book offering.
b. Based on the response models, compute a score for each customer in the database. Use this score and a cut-off value to extract a target customer list for direct mail promotion.

Targeting promotions was considered to be of prime importance. There were, in addition, other opportunities to create successful marketing campaigns based on customer behavior data such as returns, inactivity, complaints, and compliments. CBC planned to address these opportunities at a subsequent stage.

Art History of Florence

A new title, "The Art History of Florence", is ready for release. CBC has sent a test mailing to a random sample of 4,000 customers from its customer base. The customer responses have been collated with past purchase data. The data should be randomly partitioned into 3 parts- **Training Data** (1800 customers): initial data to be used to fit response models, **Validation Data** (1400 customers): hold-out data used to compare the performance of different response models, and **Test Data** (800 customers): data only to be used after a final model has been selected to estimate the likely accuracy of the model when it is deployed. The Sample Data are in a spreadsheet Charlesl.xls. Each row (or case) in the spreadsheet (other than the header) corresponds to one market test customer. Each column is a variable with the header row giving the name of the variable. The variable names and descriptions are given in Table 1, below.

Table 1: List of Variables in Charles.xls

Variable Name	**Description**
Seq#	Sequence number in the partition
ID#	Identification number in the full (unpartitioned) market test data set
Gender	O=Male 1=Female
M	Monetary- Total money spent on books
R	Recency- Months since last purchase
F	Frequency - Total number of purchases
FirstPurch	Months since first purchase
ChildBks	Number of purchases from the category: Child books
YouthBks	Number of purchases from the category: Youth books
CookBks	Number of purchases from the category: Cookbooks
DoItYBks	Number of purchases from the category: Do It Yourself books I
RefBks	Number of purchases from the category: Reference books (Atlases, Encyclopedias, Dictionaries)
ArtBks	Number of purchases from the category: Art books
GeoBks	Number of purchases from the category: Geography books
ItalCook	Number of purchases of book title: "Secrets of Italian Cooking"
ItalAtlas	Number of purchases of book title: "Historical Atlas of Italy"
ItalArt	Number of purchases of book title: "Italian Art"
Florence	=1 'The Art History of Florence" was bought, = 0 if not
Related purchase	Number of related books purchased

DATA MINING TECHNIQUES

There are various data mining techniques that can be used to mine the data collected from the market test. No one technique is universally better than another. The particular context and the particular characteristics of the data are the major factors in determining which techniques perform better in an application. For homework assignments 1, 4, and 7 we will focus on two fundamental techniques:

- Logistic regression
- K-Nearest Neighbor

We will compare them with each other as well as with a standard industry practice known as RFM Segmentation.

RFM Segmentation (homework 1 – further details on assignments page of course website)

The segmentation process in database marketing aims to partition customers in a list of prospects into homogenous groups (segments) that are similar with respect to buying behavior. The homogeneity criterion we need for segmentation is propensity to purchase the offering. But since we cannot measure this attribute, we use variables that are plausible indicators of this propensity. In the direct marketing business the most commonly used variables are the 'RFM variables':

R - Recency - time since last purchase

F - Frequency - the number of previous purchases from the company over a period

M - Monetary - the amount of money spent on the company's products over a period

The assumption is that the more recent the last purchase, the more products bought from the company in the past, and the more money spent in the past buying the company's products, the more likely is the customer to purchase the product offered.

The 1800 observations in the training data and the 1400 observations in the validation data in Charles have been divided into Recency, Frequency and Monetary categories as follows:

Recency:
0-2 months (Rcode=1)
3-6 months (Rcode=2)
7-12 months (Rcode=3)
13 months and up (Rcode=4)

Frequency:
1 book (Fcode=l)
2 books (Fcode=2)
3 books and up (Fcode=3)

Monetary:
$0 - $25 (Mcode=1)
$26 - $50 (Mcode=2)
$51- $100 (Mcode=3)
$101- $200 (Mcode=4)
$201 and up (Mcode=5)

The tables below display the 1800 customers in the training data, cross-tabulated by these categories. The purchasers are summarized in the first five tables and all the customers sent offers in the next five tables.

Purchasers

Sum of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	2	2	10	7	17	38
2		3	5	9	17	34
3		1	1	15	62	79
Grand Total	2	6	16	31	96	151

Rcode	1					
Sum of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	0	0	0	2	1	3
2		1	0	0	1	2
3		1	0	0	5	6
Grand Total	0	2	0	2	7	11

Rcode	2					
Sum of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	1	0	1	1	5	8
2		0	3	5	5	13
3			0	4	10	14
Grand Total	1	0	4	10	20	35

Rcode	3					
Sum of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	1	0	1	2	5	9
2		1	1	2	4	8
3		0	0	4	31	35
Grand Total	1	1	2	8	40	52

Rcode	4					
Sum of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	0	2	8	2	6	18
2		1	1	2	7	11
3			1	7	16	24
Grand Total	0	3	10	11	29	53

All customers

Count of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	20	40	93	166	219	538
2		32	91	180	247	550
3		2	33	179	498	712
Grand Total	20	74	217	525	964	1800

Rcode	1					
Count of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	2	2	6	10	15	35
2		3	4	12	16	35
3		1	2	11	45	59
Grand Total	2	6	12	33	76	129

Rcode	2					
Count of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	3	5	17	28	26	79
2		2	17	30	31	80
3			3	34	66	103
Grand Total	3	7	37	92	123	262

Rcode	3					
Count of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	7	15	24	51	86	183
2		12	29	55	85	181
3		1	17	53	165	236
Grand Total	7	28	70	159	336	600

Rcode	4					
Count of Florence	Mcode					
Fcode	1	2	3	4	5	Grand Total
1	8	18	46	77	92	241
2		15	41	83	115	254
3			11	81	222	314
Grand Total	8	33	98	241	429	809

Logistic Regression (homework 4 – further details on assignments page of course website)

The Logistic Regression model offers a powerful method for modeling response because it yields well-defined purchase probabilities. (The model is especially attractive in consumer choice settings because it can be derived from the random utility theory of consumer behavior, under the assumption that the error term in the customer's utility function follows a type I extreme value distribution.)

Use the Training set data of 1800 observations to construct three logistic regression models with:

- The full set of 15 predictors in the data set as input variables and "Florence" as the output variable,
- a subset that you judge as the best,
- only the R, F, M variables.

k- Nearest Neighbor (homework 7 – further details on assignments page of course website)

The k-Nearest Neighbor technique can be used to create segments based on product proximity of the offered products to similar products as well as propensity to purchase (as measured by the RFM variables). For "The Art History of Florence ", a possible segmentation by product proximity could be created using the following variables:

1. M: Monetary - Total money ($) spent on books
2. R: Recency - Months since last purchase
3. F : Frequency -Total number of past purchases
4. FirstPurch : Months since first purchase
5. RelatedPurch: Total number of past purchases of related books, i.e. sum of purchases from Art and Geography categories and of titles "Secrets of Italian Cooking", "Historical Atlas of Italy", and "Italian Art".

References

[1] Association of American Publishers. Industry Statistics, 2002.