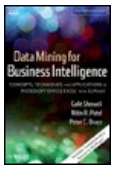

Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, Second Edition

by Galit Shmueli, Nitin R. Patel and Peter C. Bruce
John Wiley & Sons (US). (c) 2010. Copying Prohibited.

Chapter 18: Cases

18.1 Charles Book Club

CharlesBookClub.xls is the dataset for this case study.

The Book Industry

Approximately 50, 000 new titles, including new editions, are published each year in the United States, 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 United States in the 1970s was characterized by the growth of bookstore chains located in shopping malls. The 1980s saw increased purchases in bookstores stimulated through the widespread practice of discounting. By the 1990s, 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, 000-80, 000 titles and employ well-informed sales personnel. Superstores apply intense competitive pressure on book clubs and mail-order firms as well on as traditional book retailers. In response to these pressures, book clubs have 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 the continuity and negative option programs, which extended contractual relationships between the club and its members. Under a *continuity program*, a reader signs 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 book 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 is delivered to them automatically unless they specifically mark "no" on their order form by a deadline date. 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 target customers more accurately. 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 design special programs carefully tailored to meet their customer segments' varying needs.

Database Marketing at Charles

The Club The Charles Book Club (CBC) was established in December 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 created an active database of 500, 000 readers; most were acquired through advertising in specialty magazines.

The Problem CBC sent mailings to its club members each month containing the latest offerings. On the surface, CBC appeared very successful: Mailing volume was increasing, book selection was diversifying and growing, and its customer

database was increasing. However, its bottom-line profits were falling. The decreasing profits led CBC to revisit its original plan of using database marketing to improve mailing yields and to stay prof itable.

A Possible Solution CBC embraced the idea of deriving intelligence from its data to allow CBC to know its customers 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 CBC had in place were:

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

For each new title, CBC decided to use a two-step approach:

1. Conduct a market test involving a random sample of 7000 customers from the database to enable analysis of customer responses. The analysis would create and calibrate response models for the current book offering.
2. Based on the response models, compute a score for each customer in the database. Use this score and a cutoff value to extract a target customer list for direct-mail promotion.

Targeting promotions was considered to be of prime importance. Other opportunities to create successful marketing campaigns based on customer behavior data (returns, inactivity, complaints, compliments, etc.) would be addressed by CBC at a later stage.

Art History of Florence A new title, *The Art History of Florence*, is ready for release. CBC sent a test mailing to a random sample of 4000 customers from its customer base. The customer responses have been collated with past purchase data. The dataset has been randomly partitioned into three parts: *Training Data* (1800 customers)— initial data to be used to fit response models; *Validation Data* (1400 customers)— holdout data used to compare the performance of different response models; and *Test Data* (800 customers)— data to be used only after a final model has been selected to estimate the probable performance of the model when it is deployed. 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 18.1.

Table 18.1: LIST OF VARIABLES IN CHARLES BOOK CLUB DATASET

Variable Name	Description
Seq#	Sequence number in the partition
ID#	Identification number in the full (unpartitioned) market test dataset
Gender	0 = 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
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; = 0ifnot
Related Purchase	Number of related books purchased

Data Mining Techniques

Various data mining techniques 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 this assignment we focus on two fundamental techniques: fe-nearest neighbors and logistic regression. We compare them with each other as well as with a standard industry practice known as *RFM (recency, frequency, monetary) segmentation*.

RFM Segmentation The segmentation process in database marketing aims to partition customers in a list of prospects into homogeneous groups (segments) that are similar with respect to buying behavior. The homogeneity criterion we need for segmentation is the 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, number of previous purchases from the company over a period

M monetary, 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 the customer is to purchase the product offered.

The 1800 observations in the training data and the 1400 observations in the validation data 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)

Tables 18.2 and 18.3 display the 1800 customers in the training data cross-tabulated by these categories. Both buyers and nonbuyers are summarized. These tables are available for Excel computations in the RFM spreadsheet in the data file.

Table 18.2: RFM COUNTS FOR BUYERS

Rcode=all 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=i 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	**i**	**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

Table 18.3: RFM COUNTS FOR ALL CUSTOMERS (BUYERS AND NONBUYERS)

Rcode = all 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

Assignment

1. What is the response rate for the training data customers taken as a whole? What is the response rate for each of the 4 × 5 × 3 = 60 combinations of RFM categories? Which combinations have response rates in the training data that are above the overall response in the training data?

2. Suppose that we decide to send promotional mail only to the "above-average" RFM combinations identified in part 1. Compute the response rate in the validation data using these combinations.

3. Rework parts 1 and 2 with three segments:

   Segment 1: Consisting of RFM combinations that have response rates that exceed twice the overall response rate

   Segment 2: Consisting of RFM combinations that exceed the overall response rate but do not exceed twice that rate

   Segment 3: Consisting of the remaining RFM combinations

   Draw the cumulative lift curve (consisting of three points for these three segments) showing the number of customers in the validation dataset on the *x* axis and cumulative number of buyers in the validation dataset on the *y* axis.

The κ-Nearest Neighbors The κ-nearest neighbor technique can be used to create segments based on product proximity to similar products of the products offered as well as the 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:

M: Monetary—total money (in dollars) spent on books

R: Recency—months since last purchase

F: Frequency—total number of past purchases

FirstPurch: Months since first purchase

RelatedPurch: Total number of past purchases of related books (i.e., sum of purchases from the art and geography categories and of titles *Secrets of Italian Cooking, Historical Atlas of Italy*, and *Italian Art)*.

4. Use the fe-nearest neighbor option under the *Classify* menu choice in XLMiner to classify cases with $\kappa = 1$, $\kappa = 3$, and $\kappa = 11$. Use normalized data (note the checkbox "normalize input data" in the dialog box) and all five variables.
5. Use the κ-nearest neighbor option under the *Prediction* menu choice in XLMiner to compute a cumulative gains curve for the validation data for $\kappa = 1$, $\kappa = 3$, and $\kappa = 11$. Use normalized data (note the checkbox "normalize input data" in the dialog box) and all five variables. The κNN prediction algorithm gives a numerical value, which is a weighted average of the values of the Florence variable for the κ-nearest neighbors with weights that are inversely proportional to distance.

   Logistic Regression 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 dataset as independent variables and Florence as the dependent variable
   - A subset that you judge to be the best
   - Only the *R, F*, and *M* variables
6. Score the customers in the validation sample and arrange them in descending order of purchase probabilities.
7. Create a cumulative gains chart summarizing the results from the three logistic regression models created above, along with the expected cumulative gains for a random selection of an equal number of customers from the validation dataset.
8. If the cutoff criterion for a campaign is a 30% likelihood of a purchase, find the customers in the validation data that would be targeted and count the number of buyers in this set.

[1]The Charles Book Club case was derived, with the assistance of Ms. Vinni Bhandari, from *The Bookbinders Club, a Case Study in Database Marketing*, prepared by Nissan Levin and Jacob Zahavi, Tel Aviv University; used with permission.

18.2 German Credit

GermanCredit.xls is the dataset for this case study.

The German Credit dataset[2] has 30 variables and 1000 records, each record being a prior applicant for credit. Each applicant was rated as "good credit" (700 cases) or "bad credit" (300 cases).

New applicants for credit can also be evaluated on these 30 predictor variables and classified as a good or a bad credit risk based on the predictor variables. All the variables are explained in Table 18.4.

Note The original dataset had a number of categorical variables, some of which have been transformed into a series of binary variables so that they can be handled appropriately by XLMiner. Several ordered categorical variables have been left as is, to be treated by XLMiner as numerical.

Table 18.4: VARIABLES FOR THE GERMAN CREDIT DATASET

Var.	Variable Name	Description	Variable Type	Code Description
1	OBS#	Observation numbers	Categorical	Sequence number in dataset
2	CHK_ACCT	Checking account status	Categorical	0:< 0DM

				1:0 -<= … < 200 DM 2: => 200 DM 3: no checking account
3	DURATION	Duration of credit in months	Numerical	
4	HISTORY	Credit history	Categorical	0: no credits taken 1: all credits at this bank paid back duly 2: existing credits paid back duly until now 3: delay in paying off in the past 4: critical account
5	NEWCAR	Purpose of credit	Binary	car (new) 0: No, 1: Yes
6	USED_CAR	Purpose of credit	Binary	car (used) 0: No, 1: Yes
7	FURNITURE	Purpose of credit	Binary	furniture/equipment 0: No, 1: Yes
8	RADIO/TV	Purpose of credit	Binary	radio/television 0: No, 1: Yes
9	EDUCATION	Purpose of credit	Binary	education 0: No, 1: Yes
10	RETRAINING	Purpose of credit	Binary	retraining 0: No, 1: Yes
11	AMOUNT	Credit amount	Numerical	
12	SAV_ACCT	Average balance in savings account	Categorical	0:< 100 DM 1: 100 <= … < 500 DM 2: 500 <= … < 1000 DM 3: => 1000 DM 4: unknown/ no savings account
13	EMPLOYMENT	Present employment since	Categorical	0: unemployedy > 1: < 1 year 2: 1 <= … < 4 years 3:4 <= … < 7 years 4: >= 7 years
14	INSTALLRATE	Installment rate as%of disposable income	Numerical	
15	MALEJDIV	Applicant is male and divorced	Binary	0: No, 1:Yes
16	MALE_SINGLE	Applicant is male and single	Binary	0: No, 1:Yes
17	MALE_MAR_WID	Applicant is male and marriedor a widower	Binary	0: No, 1:Yes
18	CO-APPLICANT	Application has a coapplicant	Binary	0: No, 1:Yes
19	GUARANTOR	Applicant has a guarantor	Binary	0: No, 1:Yes
20	PRESENT_RESIDENT	Present resident since (years)	Categorical	0:<= 1 year 1 < … <= 2 years 2 < … <= 3 years 3:> 4 years
21	REAL_ESTATE	Applicant owns real estate	Binary	0: No, 1:Yes
22	PROPJJNKNJNONE	Applicant owns no property (or unknown)	Binary	0: No, 1:Yes
23	AGE	Age in years	Numerical	
24	OTHERJNSTALL	Applicant has other installment plan credit	Binary	0: No, 1:Yes
25	RENT	Applicant rents	Binary	0: No, 1:Yes
26	OWNLRES	Applicant owns residence	Binary	0: No, 1:Yes
27	NUIVLCREDITS	Number of existing credits at this bank	Numerical	

28	JOB	Nature of job	Categorical	0: unemployed/ unskilled— non-resident 1: unskilled— resident 2: skilled employee/ official 3: management/ self-employed/ highly qualified employee/officer
29	NUIVLDEPENDENTS	Number of people for whom liable to provide maintenance	Numerical	
30	TELEPHONE	Applicant has phone in his or her name	Binary	0: No, 1:Yes
31	FOREIGN	Foreign worker	Binary	0: No, 1:Yes
32	RESPONSE	Good credit rating	Binary	0: No, 1:Yes

Figure 18.1 shows the values of these variables for the first several records in the case.

OBS#	CHK_ACCT	DURATION	HISTORY	NEW_CAR	USED_CAR	FURNITURE	RADIO/TV	EDUCATION	RETRAINING	AMOUNT	SAV_ACCT	EMPLOYMENT	INSTALL_RATE	MALE_DIV
1	0	6	4	0	0	0	1	0	0	1169	4	4	4	0
2	1	48	2	0	0	0	1	0	0	5951	0	2	2	0
3	3	12	4	0	0	0	0	1	0	2096	0	3	2	0
4	0	42	2	0	0	1	0	0	0	7882	0	3	2	0

MALE_MAR_or_WID	CO-APPLICANT	GUARANTOR	PRESENT_RESIDENT	REAL_ESTATE	PROP_UNKN_NONE	AGE	OTHER_INSTALL	RENT	OWN_RES	NUM_CREDITS	JOB	NUM_DEPENDENTS	TELEPHONE	FOREIGN
0	0	0	4	1	0	67	0	0	1	2	2	1	1	0
0	0	0	2	1	0	22	0	0	1	1	2	1	0	0
0	0	0	3	1	0	49	0	0	1	1	1	2	0	0
0	0	1	4	0	0	45	0	0	0	1	2	2	0	0

FIGURE 18.1: DATA SAMPLE (FIRST SEVERAL ROWS)

The consequences of misclassification have been assessed as follows: The costs of a false positive (incorrectly saying that an applicant is a good credit risk) outweigh the benefits of a true positive (correctly saying that an applicant is a good credit risk) by a factor of 5. This is summarized in Table 18.5. The opportunity cost table was derived from the average net profit per loan as shown in Table 18.6.

Table 18.5: OPPORTUNITY COST TABLE (DEUTSCHE MARKS)

	Predicted (Decision)	
Actual	**Good (Accept)**	**Bad (Reject)**
Good	0	100
Bad	500	0

Table 18.6: AVERAGE NET PROFIT (DEUTSCHE MARKS)

	Predicted (Decision)	
Actual	**Good (Accept)**	**Bad (Reject)**
Good	100	0
Bad	-500	0

Because decision makers are used to thinking of their decision in terms of net profits, we use these tables in assessing the performance of the various models.

Assignment

1. Review the predictor variables and guess what their role in a credit decision might be. Are there any surprises in the data?
2. Divide the data into training and validation partitions, and develop classification models using the following data mining techniques in XLMiner: logistic regression, classification trees, and neural networks.
3. Choose one model from each technique and report the confusion matrix and the cost/gain matrix for the validation data. Which technique has the most net profit?
4. Let us try and improve our performance. Rather than accept XLMiner's initial classification of all applicants' credit status, use the "predicted probability of success" in logistic regression (where *success* means 1) as a basis for selecting the best credit risks first, followed by poorer risk applicants.
    a. Sort the validation on "predicted probability of success."
    b. For each case, calculate the net profit of extending credit.
    c. Add another column for cumulative net profit.
    d. How far into the validation data do you go to get maximum net profit? (Often, this is specified as a percentile or rounded to deciles.)
    e. If this logistic regression model is scored to future applicants, what "probability of success" cutoff should be used in extending credit?

[2]This is available from ftp.ics.uci.edu/pub/machine-learning-databases/statlog.

18.3 Tayko Software Cataloger

Tayko.xls is the dataset for this case study.

Background

Tayko is a software catalog firm that sells games and educational software.[3] It started out as a software manufacturer and later added third-party titles to its offerings. It has recently put together a revised collection of items in a new catalog, which it is preparing to roll out in a mailing.

In addition to its own software titles, Tayko's customer list is a key asset. In an attempt to expand its customer base, it has recently joined a consortium of catalog firms that specialize in computer and software products. The consortium affords members the opportunity to mail catalogs to names drawn from a pooled list of customers. Members supply their own customer lists to the pool and can "withdraw" an equivalent number of names each quarter. Members are allowed to do predictive modeling on the records in the pool so they can do a better job of selecting names from the pool.

The Mailing Experiment

Tayko has supplied its customer list of 200, 000 names to the pool, which totals over 5 million names, so it is now entitled to draw 200, 000 names for a mailing. Tayko would like to select the names that have the best chance of performing well, so it conducts a test—it draws 20, 000 names from the pool and does a test mailing of the new catalog.

This mailing yielded 1065 purchasers, a response rate of 0.053. Average spending was $103 for each of the purchasers, or $5.46 per catalog mailed. To optimize the performance of the data mining techniques, it was decided to work with a stratified sample that contained equal numbers of purchasers and nonpurchasers. For ease of presentation, the dataset for this case includes just 1000 purchasers and 1000 nonpurchasers, an apparent response rate of 0.5. Therefore, after using the dataset to predict who will be a purchaser, we must adjust the purchase rate back down by multiplying each case's "probability of purchase" by 0.053/0.5, or 0.107.

Data

There are two response variables in this case. *Purchase* indicates whether or not a prospect responded to the test mailing and purchased something. *Spending* indicates, for those who made a purchase, how much they spent. The overall

procedure in this case will be to develop two models. One will be used to classify records as Purchase or No purchase. The second will be used for those cases that are classified as *purchase* and will predict the amount they will spend.

Table 18.7 provides a description of the variables available in this case. A partition variable is used because we will be developing two different models in this case and want to preserve the same partition structure for assessing each model.

Table 18.7: DESCRIPTION OF VARIABLES FOR TAYKO DATASET

Var.	Variable Name	Description	Variable Type	Code Description
1	US	Is it a U.S. address?	Binary	1: yes
				0:no
2-16	Source_*	Source catalog for the record (15 possible sources)	Binary	1: yes
				0:no
17	Freq.	Number of transactions in last year at source catalog	Numerical	
18	last_update_days_ago	How many days ago last update was made to customer record	Numerical	
19	1st_update_days_ago	How many days ago first update was made to customer record	Numerical	
20	RFM%	Recency-frequency-monetary percentile, as reported by source catalog (see Section 18.1)	Numerical	
21	Web_order	Customer placed at least one order via Web	Binary	1: yes
				0:no
22	Gender=mal	Customer is male	Binary	1: yes
				0:no
23	Address_is_res	Address is a residence	Binary	1: yes
				0:no
24	Purchase	Person made purchase in test mailing	Binary	1: yes
				0:no
25	Spending	Amount (dollars) spent by customer in test mailing	Numerical	
26	Partition	Variable indicating which partition the record will be assigned to	Alphabetical	t: trainingv: validations: test

Figure 18.2 shows the first few rows of data (the top shows the sequence number plus the first 14 variables, and the bottom shows the remaining 11 variables for the same rows).

sequence_number	US	source_a	source_c	source_b	source_d	source_e	source_m	source_o	source_h	source_r	source_s	source_t	source_u	source_p
1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	1	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	1	0	0
4	1	0	1	0	0	0	0	0	0	0	0	0	0	0
5	1	0	1	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	1	0	0	0	0
7	1	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	0	0	1	0	0	0	0	0	0	0	0	0	0
9	1	1	0	0	0	0	0	0	0	0	0	0	0	0
10	1	1	0	0	0	0	0	0	0	0	0	0	0	0

source_x	source_w	Freq	last_update_days_ago	1st_update_days_ago	Web order	Gender=male	Address_is_res	Purchase	Spending	Partition
0	0	2	3662	3662	1	0	1	1	128	s
0	0	0	2900	2900	1	1	0	0	0	s
0	0	2	3883	3914	0	0	0	1	127	t
0	0	1	829	829	0	1	0	0	0	s
0	0	1	869	869	0	0	0	0	0	t
0	0	1	1995	2002	0	0	1	0	0	s
0	1	2	1498	1529	0	0	1	0	0	s
0	0	1	3397	3397	0	1	0	0	0	t
0	0	4	525	2914	1	1	0	1	489	t
0	0	1	3215	3215	0	0	0	1	174	v

FIGURE 18.2: DATA FOR FIRST 10 RECORDS

Assignment

1. Each catalog costs approximately $2 to mail (including printing, postage, and mailing costs). Estimate the gross profit that the firm could expect from the remaining 180, 000 names ifit selected them randomly from the pool.
2. Develop a model for classification of a customer as a purchaser or non-purchaser.
    a. Partition the data into training data on the basis of the partition variable, which has 800 *t*'s, 700 *v*'s, and 500 *s*'s (training data, validation data, and test data, respectively) assigned randomly to cases.
    b. Using the "best subset" option in logistic regression, implement the full logistic regression model, select the best subset of variables, and then implement a regression model with just those variables to classify the data into purchasers and nonpurchasers. (Logistic regression is used because it yields an estimated "probability of purchase," which is required later in the analysis.)
3. Develop a model for predicting spending among the purchasers.
    a. Make a copy of the data sheet (call it data2), sort by the Purchase variable, and remove the records where Purchase = 0 (the resulting spreadsheet will contain only purchasers).
    b. Partition this dataset into training and validation partitions on the basis of the partition variable.
    c. Develop models for predicting spending, using:
        i. Multiple linear regression (use best subset selection)
        ii. Regression trees
    d. Choose one model on the basis of its performance with the validation data.
4. Return to the original test data partition. Note that this test data partition includes both purchasers and nonpurchasers. Note also that although it contains the scoring of the chosen classification model, we have not used this partition in our analysis up to this point, so it will give an unbiased estimate of the performance of our models. It is best to make a copy of the test data portion of this sheet to work with, since we will be adding analysis to it. This copy is called *Score Analysis*.
    a. Copy to this sheet the "predicted probability of success" (Success = Purchase) column from the classification of test data.

b. Score to this data sheet the prediction model chosen.

c. Arrange the following columns so that they are adjacent:
   i. Predicted probability of purchase (Success)
   ii. Actual spending (dollars)
   iii. Predicted spending (dollars)

d. Add a column for "adjusted probability of purchase" by multiplying "predicted probability of purchase" by 0.107. *This is to adjust for over-sampling the purchasers* (see above).

e. Add a column for expected spending (adjusted probability of purchase × predicted spending).

f. Sort all records on the "expected spending" column.

g. Calculate cumulative lift [= cumulative "actual spending" divided by the average spending that would result from random selection (each adjusted by 0.107)].

h. Using this cumulative lift curve, estimate the gross profit that would result from mailing to the 180, 000 names on the basis of your data mining models.

Note Although Tayko is a hypothetical company, the data in this case (modified slightly for illustrative purposes) were supplied by a real company that sells software through direct sales. The concept of a catalog consortium is based on the Abacus Catalog Alliance. Details can be found at www.doubleclick.com/us/solutions/marketers/database/catalog/.

[3]Resampling Stats, Inc. 2006; used with permission.

18.4 Segmenting Consumers of Bath Soap

BathSoap.xls is the dataset for this case study.

Business Situation

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in consumer goods (both durable and nondurable).[4] In one major research project, CRISA tracks numerous consumer product categories (e.g. "detergents") and, within each category, perhaps dozens of brands. To track purchase behavior, CRISA constituted household panels in over 100 cities and towns in India, covering most of the Indian urban market. The households were carefully selected using stratified sampling to ensure a representative sample; a subset of 600 records is analyzed here. The strata were defined on the basis of socioeconomic status and the market (a collection of cities).

CRISA has both transaction data (each row is a transaction) and household data (each row is a household), and for the household data, maintains the following information:

- Demographics of the households (updated annually)
- Possession of durable goods (car, washing machine, etc., updated annually; an "affluence index" is computed from this information)
- Purchase data of product categories and brands (updated monthly)

CRISA has two categories of clients: (1) advertising agencies that subscribe to the database services, obtain updated data every month, and use the data to advise their clients on advertising and promotion strategies; (2) and consumer goods manufacturers, which monitor their market share using the CRISA database.

Key Problems

CRISA has traditionally segmented markets on the basis of purchaser demographics. It would now like to segment the market based on two key sets of variables more directly related to the purchase process and to brand loyalty:

1. Purchase behavior (volume, frequency, susceptibility to discounts, and brand loyalty)

2. Basis of purchase (price, selling proposition)

Doing so would allow CRISA to gain information about what demographic attributes are associated with different purchase behaviors and degrees of brand loyalty, and thus deploy promotion budgets more effectively. More effective market segmentation would enable CRISA's clients (in this case, a firm called IMRB) to design more cost-effective promotions targeted at appropriate segments. Thus, multiple promotions could be launched, each targeted at different market segments at different times of the year. This would result in a more cost-effective allocation of the promotion budget to different market segments. It would also enable IMRB to design more effective customer reward systems and thereby increase brand loyalty.

Data

The data in Table 18.8 profile each household, each row containing the data for one household.

Table 18.8: DESCRIPTION OF VARIABLES FOR EACH HOUSEHOLD

Variable Type	Variable Name	Description
Member ID Demographics	Member id SEC FEH	Unique identifier for each household Socioeconomic class (1 = high, 5 = low) Eating habits (1 =vegetarian, 2 = vegetarian but eat eggs, 3 =nonvegetarian, 0 = not specified)
	MT SEX	Native language (see table in worksheet) Gender of homemaker (1=male, 2 = female)
	AGE EDU	Age of homemaker Education of homemaker (1 = minimum, 9 = maximum)
	HS CHILD	Number of members in household Presence of children in household (4 categories)
	CS	Television availability (1 = available, 2 = unavailable)
	Affluence Index	Weighted value of durables possessed
Purchase summary over the period	No. of Brands	Number of brands purchased
	Brand Runs	Number of instances of consecutive purchase of brands
	Total Volume	Sum of volume
	No. of Trans	Number of purchase transactions (multiple brands purchased in a month are counted as separate transactions
	Value	Sum of value
	Trans/ Brand Runs	Average transactions per brand run
	Vol/Trans	Average volume per transaction
	Avg. Price	Average price of purchase
Purchase within promotion	Pur Vol	Percent of volume purchased
	No Promo - %	Percent of volume purchased under no promotion
	Pur Vol Promo 6%	Percent of volume purchased under promotion code 6
	Pur Vol Other Promo %	Percent of volume purchased under other promotions
Brandwise purchase	Br. Cd. (57, 144), 55, 272, 286, 24, 481, 352, 5, and 999 (others)	Percent of volume purchased of the brand
Price categorywise purchase	Price Cat 1 to 4	Percent of volume purchased under the price category
Selling propositionwise purchase	Proposition Cat 5 to 15	Percent of volume purchased under the product proposition category

Measuring Brand Loyalty

Several variables in this case measure aspects of brand loyalty. The number of different brands purchased by the

customer is one measure. However, a consumer who purchases one or two brands in quick succession, then settles on a third for a long streak, is different from a consumer who constantly switches back and forth among three brands. How often customers switch from one brand to another is another measure of loyalty. Yet a third perspective on the same issue is the proportion of purchases that go to different brands—a consumer who spends 90% of his or her purchase money on one brand is more loyal than a consumer who spends more equally among several brands.

All three of these components can be measured with the data in the purchase summary worksheet.

Assignment

1. Use κ-means clustering to identify clusters of households based on:
    a. The variables that describe purchase behavior (including brand loyalty)
    b. The variables that describe the basis for purchase
    c. The variables that describe both purchase behavior and basis of purchase

    Note 1: How should κ be chosen? Think about how the clusters would be used. It is likely that the marketing efforts would support two to five different promotional approaches.

    Note 2: How should the percentages of total purchases comprised by various brands be treated? Isn't a customer who buys all brand A just as loyal as a customer who buys all brand B? What will be the effect on any distance measure of using the brand share variables as is? Consider using a single derived variable.
2. Select what you think is the best segmentation and comment on the characteristics (demographic, brand loyalty, and basis for purchase) of these clusters. (This information would be used to guide the development of advertising and promotional campaigns.)
3. Develop a model that classifies the data into these segments. Since this information would most likely be used in targeting direct-mail promotions, it would be useful to select a market segment that would be defined as a *success* in the classification model.

Appendix

Although not used in the assignment, two additional datasets were used in the derivation of the summary data.

CRISA_Purchase_Data is a transaction database in which each row is a transaction. Multiple rows in this dataset corresponding to a single household were consolidated into a single household row in CRISA_Summary_Data.

The *Durables* sheet in IMRB_Summary_Data contains information used to calculate the affluence index. Each row is a household, and each column represents a durable consumer good. A 1 in the column indicates that the durable is possessed by the household; a 0 indicates that it is not possessed. This value is multiplied by the weight assigned to the durable item. For example, a 5 indicates the weighted value of possessing the durable. The sum of all the weighted values of the durables possessed equals the affluence index.

[4]Cytel, Inc. and Resampling Stats, Inc. 2006; used with permission.

18.5 Direct-Mail Fundraising

Fundraising.xls and FutureFundraising.xls are the datasets used for this case study.

Background

A national veterans' organization wishes to develop a data mining model to improve the cost-effectiveness of its direct marketing campaign. The organization, with its in-house database of over 13 million donors, is one of the largest direct-mail fundraisers in the United States. According to its recent mailing records, the overall response rate is 5.1%. Out of those who responded (donated), the average donation is $13.00. Each mailing, which includes a gift of personalized address labels and assortments of cards and envelopes, costs $0.68 to produce and send. Using these facts, we take a sample of this dataset to develop a classification model that can effectively capture donors so that the expected net profit is maximized. Weighted sampling is used, underrepresenting the nonresponders so that the sample has equal numbers of donors and nondonors.

Data

The file Fundraising.xls contains 3120 data points with 50% donors (TARGET-B = 1) and 50% nondonors (TARGET_B = 0). The amount of donation (TARGET_D) is also included but is not used in this case. The descriptions for the 25 variables (including 2 target variables) are listed in Table 18.9.

Table 18.9: DESCRIPTION OF VARIABLES FOR THE FUNDRAISING DATASET

Variable	Description
ZIP	Zip code group (Zip codes were grouped into five groups; only four are needed for analysis, since if a potential donor falls into none of the four, s/he must be in the other group. Inclusion of all five variables is redundant and will cause some methods to fail. "1" indicates that the potential donor belongs to this zip group. 00000-19999 ⇒ 1 (omitted for reason stated above) 20000-39999 ⇒ zipconvert-2 40000-59999 ⇒ zipconvert-3 60000-79999 ⇒ zipconvert-4 80000-99999 ⇒ zipconvert-5
HOMEOWNER	1 = homeowner, 0 = not a homeowner
NUMCHLD	Number of children
INCOME	Household income
GENDER	0 = male, 1 = female
WEALTH	Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. Segments are denoted 0-9 (0=lowest wealth group, 9 = highest wealth group). Each rating has a different meaning within each state.
HV	Average home value in potential donor's neighborhood in hundreds of dollars
ICmed	Median family income in potential donor's neighborhood in hundreds of dollars
ICavg	Average family income in potential donor's neighborhood in hundreds
IC15	Percent earning less than $15K in potential donor's neighborhood
NUMPROM	Lifetime number of promotions received to date
RAMNTALL	Dollar amount of lifetime gifts to date
MAXRAMNT	Dollar amount of largest gift to date
LASTGIFT	Dollar amount of most recent gift
TOTALMONTHS	Number of months from last donation to July 1998 (the last time the case was updated)
TIMELAG	Number of months between first and second gift
AVGGIFT	Average dollar amount of gifts to date
TARGET_*B*	Target variable: binary indicator for response (1 = donor, 0 = nondonor)
TARGET_*D*	Target variable: donation amount (in dollars). We will NOT use this variable for this case.

Assignment

Step 1: Partitioning. Partition the dataset into 60% training and 40% validation (set the seed to 12345).

Step 2: Model Building. Follow these steps:

1. *Selecting classification tool and parameters*. Run the following classification tools on the data:
   - Logistic regression
   - Classification trees
   - Neural networks

   Be sure to test different parameter values for each method. You may also want to run each method on a subset of the

variables. Be sure NOT to include TARGET_D in your analysis.

2. *Classification under asymmetric response and cost*. What is the reasoning behind using weighted sampling to produce a training set with equal numbers of donors and nondonors? Why not use a simple random sample from the original dataset? *(Hint:* Given the actual response rate of 5.1%, how do you think the classification models will behave under simple sampling?) In this case, is classification accuracy a good performance metric for our purposes of maximizing net profit? If not, how would you determine the best model? Explain your reasoning.

3. *Calculate net profit*. For each method, calculate the lift of net profit for both the training and validation set based on the actual response rate (5.1%). Again, the expected donation, given that they are donors, is $13.00, and the total cost of each mailing is $0.68. *(Hint:* To calculate estimated net profit, we will need to undo the effects of the weighted sampling and calculate the net profit that would reflect the actual response distribution of 5.1% donors and 94.9% nondonors.)

4. *Draw lift curves*. Draw each model's net profit lift curve for the validation set onto a single graph. Do any models dominate?

5. *Best model*. From your answer in 2, what do you think is the "best" model?

Step 3: Testing. The file FutureFundraising.xls contains the attributes for future mailing candidates. Using your "best" model from step 2 (number 5), which of these candidates do you predict as donors and nondonors? List them in descending order of the probability of being a donor.

18.6 Catalog Cross Selling

CatalogCrossSell.xls is the dataset for this case study.

Background

Exeter, Inc., is a catalog firm that sells products in a number of different catalogs that it owns.[5] The catalogs number in the dozens, but fall into nine basic categories:

1. Clothing
2. Housewares
3. Health
4. Automotive
5. Personal electronics
6. Computers
7. Garden
8. Novelty gift
9. Jewelry

The costs of printing and distributing catalogs are high. By far the biggest cost of operation is the cost of promoting products to people who buy nothing. Having invested so much in the production of artwork and printing of catalogs, Exeter wants to take every opportunity to use them effectively. One such opportunity is in cross selling—once a customer has "taken the bait" and purchases one product, try to sell him or her another product while you have his or her attention.

Such cross promotion might take the form of enclosing a catalog in the shipment of a purchased product, together with a discount coupon to induce a purchase from that catalog. Or it might take the form of a similar coupon sent by e-mail, with a link to the Web version of that catalog.

But which catalog should be enclosed in the box or included as a link in the e-mail with the discount coupon? Exeter would like it to be an informed choice—a catalog that has a higher probability of inducing a purchase than simply choosing a catalog at random.

Assignment

Using the dataset CatalogCrossSell.xls, perform an association rules analysis and comment on the results. Your discussion should provide interpretations in English of the meanings of the various output statistics (lift ratio, confidence, support) and include a very rough estimate (precise calculations not necessary) of the extent to which this will help Exeter make an informed choice about which catalog to cross promote to a purchaser.

Acknowledgment The data for this case have been adapted from the data in a set of cases provided for educational purposes by the Direct Marketing Education Foundation ("DMEF Academic Data Set Two, Multi Division Catalog Company, Code: 02DMEF"); used with permission.

[5]Resampling Stats, Inc. 2006; used with permission.

18.7 Predicting Bankruptcy

Bankruptcy.xls is the dataset for this case study.

Predicting Corporate Bankruptcy[6]

Just as doctors check blood pressure and pulse rate as vital indicators of the health of a patient, so business analysts scour the financial statements of a corporation to monitor its financial health. Whereas blood pressure, pulse rate, and most medical vital signs, however, are measured through precisely defined procedures, financial variables are recorded under much less specific general principles of accounting. A primary issue in financial analysis, then, is how predictable is the health of a company?

One difficulty in analyzing financial report information is the lack of disclosure of actual cash receipts and disbursements. Users of financial statements have had to rely on proxies for cash flow, perhaps the simplest of which is income (INC) or earnings per share. Attempts to improve INC as a proxy for cash flow include using income plus depreciation (INCDEP), working capital from operations (WCFO), and cash flow from operations (CFFO). CFFO is obtained by adjusting income from operations for all noncash expenditures and revenues and for changes in the current asset and current liabilities accounts.

A further difficulty in interpreting historical financial disclosure information is caused whenever major changes are made in accounting standards. For example, the Financial Accounting Standards Board issued several promulgations in the middle 1970s that changed the requirements for reporting accruals pertaining to such things as equity earnings, foreign currency gain and losses, and deferred taxes. One effect of changes of this sort was that earnings figures became less reliable indicators of cash flow.

In the light of these difficulties in interpreting accounting information, just what are the important vital signs of corporate health? Is cash flow an important signal? If not, what is? If so, what is the best way to approximate cash flow? How can we predict the impending demise of a company?

To begin to answer some of these important questions, we conducted a study of the financial vital signs of bankrupt and healthy companies. We first identified 66 failed firms from a list provided by Dun and Bradstreet. These firms were in manufacturing or retailing and had financial data available on the Compustat Research tape. Bankruptcy occurred somewhere between 1970 and 1982.

For each of these 66 failed firms, we selected a healthy firm of approximately the same size (as measured by the book value of the firm's assets) from the same industry (3 digit SIC code) as a basis of comparison. This matched sample technique was used to minimize the impact of any extraneous factors (such as industry) on the conclusions of the study.

The study was designed to see how well bankruptcy can be predicted two years in advance. A total of 24 financial ratios were computed for each of the 132 firms using data from the Compustat tapes and from Moody's Industrial Manual for the

year that was two years prior to the year of bankruptcy. Table 18.10 lists the 24 ratios together with an explanation of the abbreviations used for the fundamental financial variables. All these variables are contained in a firm's annual report with the exception of CFFO. Ratios were used to facilitate comparisons across firms of various sizes.

Table 18.10: PREDICTING CORPORATE BANKRUPTCY: FINANCIAL VARIABLES AND RATIOS

Abbreviation	Financial Variable	Ratio	Definition
ASSETS	Total assets	R1	CASH/CURDEBT
CASH	Cash	R2	CASH/SALES
CFFO	Cash flow from operations	R3	CASH/ASSETS
COGS	Cost of goods sold	R4	CASH/DEBTS
CURASS	Current assets	R5	CFFO/SALES
CURDEBT	Current debt	R6	CFFO/ASSETS
DEBTS	Total debt	R7	CFFO/DEBTS
INC	Income	R8	COGS/INV
INCDEP	Income plus depreciation	R9	CURASS/CURDEBT
INV	Inventory	R10	CURASS/SALES
REC	Receivables	R11	CURASS/ASSETS
SALES	Sales	R12	CURDEBT/DEBTS
WCFO	Working capital from operations	R13	INC/SALES
		R14	INC/ASSETS
		R15	INC/DEBTS
		R16	INCDEP/SALES
		R17	INCDEP/ASSETS
		R18	INCDEP/DEBTS
		R19	SALES/REC
		R20	SALES/ASSETS
		R21	ASSETS/DEBTS
		R22	WCFO/SALES
		R23	WCFO/ASSETS
		R24	WCFO/DEBTS

The first four ratios using CASH in the numerator might be thought of as measures of a firm's cash reservoir with which to pay debts. The three ratios with CURASS in the numerator capture the firm's generation of current assets with which to pay debts. Two ratios, CURDEBT/DEBT and ASSETS/DEBTS, measure the firm's debt structure. Inventory and receivables turnover are measured by COGS/INV and SALES/REC, and SALES/ASSETS measures the firm's ability to generate sales. The final 12 ratios are asset flow measures.

Assignment

1. What data mining technique(s) would be appropriate in assessing whether there are groups of variables that convey the same information and how important that information is? Conduct such an analysis.

2. Comment on the distinct goals of profiling the characteristics of bankrupt firms versus simply predicting (black box style) whether a firm will go bankrupt and whether both goals, or only one, might be useful. Also comment on the classification methods that would be appropriate in each circumstance.

3. Explore the data to gain a preliminary understanding of which variables might be important in distinguishing bankrupt from nonbankrupt firms. *(Hint:* As part of this analysis, use XLMiner's boxplot option, specifying the bankrupt/not bankrupt variable as the *x* variable.)

4. Using your choice of classifiers, use XLMiner to produce several models to predict whether or not a firm goes

bankrupt, assessing model performance on a validation partition.

5. Based on the above, comment on which variables are important in classification, and discuss their effect.

[6]This case was prepared by Professor Mark E. Haskins and Professor Phillip E. Pfeifer. It was written as a basis for class discussion rather than to illustrate effective or ineffective handling of an administrative situation.

18.8 Time Series Case: Forecasting Public Transportation Demand

The dataset bicup2006.xls is used for this case study.

Background

Forecasting transportation demand is important for multiple purposes such as staffing, planning, and inventory control. The public transportation system in Santiago de Chile has gone through a major effort of reconstruction. In this context, a business intelligence competition took place in October 2006, which focused on forecasting demand for public transportation. This case is based on the competition, with some modifications.

Problem Description

A public transportation company is expecting an increase demand for its services and is planning to acquire new buses and to extend its terminals. These investments require a reliable forecast of future demand. To create such forecasts, one can use data on historic demand. The company's data warehouse has data on each 15-minute interval between 6:30 AM and 22:00, on the number of passengers arriving at the terminal. As a forecasting consultant you have been asked to create a forecasting method that can generate forecasts for the number of passengers arriving at the terminal.

Available Data

Part of the historic information is available in the file bicup2006.xls. The file contains the worksheet "Historic Information" with known demand for a 3-week period, separated into 15-minute intervals. The second worksheet ("Future") contains dates and times for a future 3-day period, for which forecasts should be generated (as part of the 2006 competition).

Assignment Goal

Your goal is to create a model/method that produces accurate forecasts. To evaluate your accuracy, partition the given historic data into two periods: a training period (the first two weeks) and a validation period (the last week). Models should be fitted only to the training data and evaluated on the validation data.

Although the competition winning criterion was the lowest Mean Absolute Error (MAE) on the future 3-day data, this is *not* the goal for this assignment. Instead, if we consider a more realistic business context, our goal is to create a model that generates reasonably good forecasts on any time/day of the week. Consider not only predictive metrics such as MAE, MAPE, and RMSE, but also look at actual and forecasted values, overlaid on a time plot.

Assignment

For your final model, present the following summary:

1. Name of the method/combination of methods.
2. A brief description of the method/combination.
3. All estimated equations associated with constructing forecasts from this method.
4. The MAPE and MAE for the training period and the validation period.
5. Forecasts for the future period (March 22-24), in 15-minute bins.
6. A single chart showing the fit of the final version of the model to the entire period (including training, validation, and

future). Note that this model should be fitted using the combined training + validation data.

Tips and Suggested Steps

1. Use exploratory analysis to identify the components of this time series. Is there a trend? Is there seasonality? If so, how many "seasons" are there? Are there any other visible patterns? Are the patterns global (the same throughout the series) or local?
2. Consider the frequency of the data from a practical and technical point of view. What are some options?
3. Compare the weekdays and weekends. How do they differ? Consider how these differences can be captured by different methods.
4. Examine the series for missing values or unusual values. Think of solutions.
5. Based on the patterns that you found in the data, which models or methods should be considered?
6. Consider how to handle actual counts of zero within the computation of MAPE.