Kenan-Flagler Business School
The University of North Carolina

Predicting Response at BookBinders: Logistic Regression (version 2)

As a direct marketer of specialty books, the BookBinders Book Club has achieved steady growth in their customer base. Yet while sales have grown steadily, profits began falling when the database got larger and when the company diversified its book selection and increased the number of offers sent to customers. The falling profits have led Stan Lawton, BookBinders' marketing director, to experiment with different database marketing approaches in order improve BookBinders' mailing yields and profits.

Stan began a series of live market tests, each involving a random sample of customers from the database. An offer for the current book selection is sent to the sample and then the sample customers' responses, either purchase or no purchase, are recorded and used to calibrate a response model for the current offering. The response model's results are then used to "score" the remaining customers in the database and select customers from the full customer database for the 'rollout' mailing campaign.

Stan's first market tests relied on RFM (recency – frequency – monetary) analysis. Direct marketers have used this approach to predict customer behavior for more than 50 years. The approach is intuitive, easy to implement, and produced significant improvements in response rates and profits compared with mass mailings to BookBinders' full database. Despite this initial success, Stan is eager to evaluate the effectiveness of alternate approaches. BookBinders offers books in different categories including cooking, art and children's' books – and the number of previous book purchases in each category is recorded in each customer's record in the database. RFM analysis does not use this or other customer information such as gender and Stan suspects that a more sophisticated modeling approach could yield superior results to the RFM approach.

Logistic Regression offers a powerful method for modeling response. Logistic regression is similar to linear regression – the key difference is that the dependent variable is binary (for example, purchase or no purchase) rather than continuous. For each customer, logistic regression predicts a probability, between 0 and 1, of purchase or response, which can be used for targeting and prediction decisions. Like linear regression, it can accommodate both

Professor Charlotte Mason prepared this case to provide material for class discussion rather than to illustrate either effective or ineffective handling of a business situation. Names and data may have been disguised to assure confidentiality. The assistance of the Direct Marketing Educational Foundation in supplying the data used for this case is gratefully acknowledged.

continuous and categorical predictors, including interaction terms. Its use in database marketing has grown as software becomes more readily available and as familiarity with the approach grows.

Stan has just received a dataset containing the responses of a random sample of 50,000 customers to a new offering from BookBinders titled "The Art History of Florence." The table below describes the variables included in the dataset:

Variable Descriptions in Test Dataset
(sample size = 50,000)

Variable name	Type	Description
ACCTNUM	Numeric	Customer account number
GENDER	Text	Customer gender: 1=male, 0=female
STATE	Text	State where customer lives (2-character abbreviation)
ZIP	Text	ZIP code (5-digit)
ZIP3	Text	First 3 digits of ZIP code
FIRST	Numeric	Number of months since first purchase
LAST	Numeric	Number of months since most recent purchase
BOOK$	Numeric	Total dollars spent on books
NONBOOK$	Numeric	Total dollars spent on non-book products
TOTAL$	Numeric	Total dollars spent
PURCH	Numeric	Total number of books purchased
CHILD	Numeric	Total number of children's books purchased
YOUTH	Numeric	Total number of youth books purchased
COOK	Numeric	Total number of cook books purchased
DO_IT	Numeric	Total number of do-it-yourself books purchased
REFERNCE	Numeric	Total number of reference books purchased
ART	Numeric	Total number of art books purchased
GEOG	Numeric	Total number of geography books purchased
BUYER	Numeric	Did the customer buy "The Art History of Florence?" (1=yes, 0=no)

Stan is eager to assess the potential value of logistic regression as a method for predicting customer response and has asked you to complete the following analyses.

Part I: Logistic Regression

Below are the results from a logistic regression model using BUYER as the dependent variable and the following as predictor variables:

LAST
TOTAL$
GENDER
CHILD
YOUTH
COOK
DO_IT
REFERNCE
ART
GEOG

Technical Note:
PURCH is excluded from the set of predictor variables – including it will lead to perfect collinearity since *PURCH* (the number of books purchased) is equal to the sum of the number of books purchased in the 7 categories. By including the number of purchases in each category, there is no need to include the total number of purchases.

Figure 1: Logistic Regression Results

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	24122.212	.117	.258

Classification Table[a]

			Predicted		
			Bought "Art History of Florence?"		Percentage Correct
	Observed		No	Yes	
Step 1	Bought "Art History of Florence?"	No	45126	352	99.2
		Yes	3838	684	15.1
	Overall Percentage				91.6

a. The cut value is .500

Figure 1 (continued)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	TOTAL$	.001	.000	31.701	1	.000	1.001
	CHILD	-.186	.017	116.097	1	.000	.830
	YOUTH	-.113	.026	18.724	1	.000	.893
	COOK	-.270	.017	249.075	1	.000	.763
	DO_IT	-.539	.027	399.777	1	.000	.583
	REFERNCE	.235	.027	78.087	1	.000	1.265
	ART	1.156	.022	2723.273	1	.000	3.176
	GEOG	.574	.019	950.087	1	.000	1.776
	GENDER(Male)	.761	.036	452.515	1	.000	2.140
	LAST	-.095	.003	1150.401	1	.000	.910
	Constant	-2.361	.049	2293.523	1	.000	.094

[a] Variable(s) entered on step 1: TOTAL$, CHILD, YOUTH, COOK, DO_IT, REFERNCE, ART, GEOG, GENDER, LAST.

1) Summarize and interpret the results (so that a marketing manager can understand them).
   a) Which variables are significant?
   b) Which seem to be 'important'?
   c) Interpret the coefficients for Art, Cook, Last, Total$ and Gender.

Part II: Decile Analysis of Logistic Regression Results

Next each customer was assigned to a decile based on his or her predicted probability of purchase – those customers with the highest probability of purchase are in decile 1, those with the lowest probability of purchase are in decile 10. Figure 2 is a bar chart plotting response rate by decile.

Figure 2: Response Rate by Purchase Probability Decile

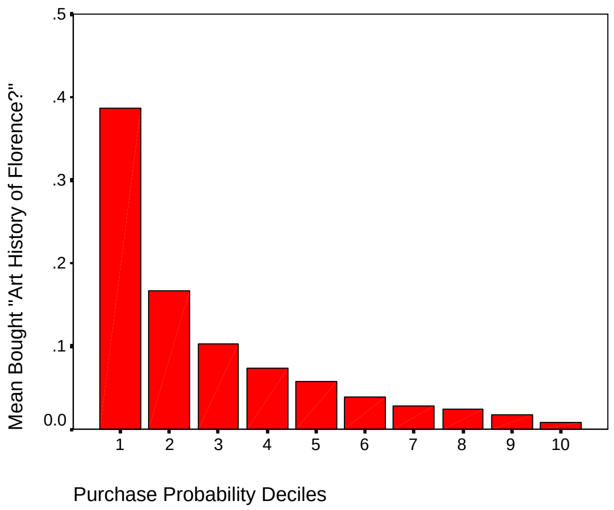

Table 1 below shows the number of customers (N), the number of buyers of “The Art History of Florence’ (Sum), and the response rate to the offer (Mean) by purchase probability decile.

Table 1: Number of Customers, Buyers and Response Rate by Purchase Probability Decile

Bought "Art History of Florence?"

Purchase Probability Deciles	N	Sum	Mean
1	5000	1935	.3870
2	5000	836	.1672
3	5000	511	.1022
4	5000	368	.0736
5	5000	284	.0568
6	5000	196	.0392
7	5001	139	.0278
8	4999	121	.0242
9	5000	90	.0180
10	5000	42	.0084
Total	50000	4522	.0904

Finally, Table 2 reports the mean values of the following variables by purchase probability decile:

Total $ spent
Months since last purchase, and
Number of books purchased for each of the seven categories (i.e., children, youth, cookbooks, do-it-yourself, reference, art and geography).

Table 2: Summary Means by Purchase Probability Decile

Mean

Purchase Probability Deciles	Total $ spent	Months since last purchase	# purchases, Children's books	# purchases, Youth books	# purchases, Cookbooks	# purchases, Do-it-yourself books	# purchases, Reference books	# purchases, Art books	# purchases, Geography books
1	257.3526	7.19	1.06	.51	1.07	.47	.56	1.50	1.33
2	224.8692	7.96	.84	.39	.85	.39	.40	.75	.89
3	214.2220	8.62	.79	.37	.80	.37	.38	.48	.70
4	207.6748	8.79	.75	.36	.80	.34	.31	.30	.54
5	199.0836	9.57	.76	.33	.82	.37	.27	.22	.46
6	199.1330	10.93	.75	.36	.86	.39	.26	.16	.39
7	191.3457	12.37	.76	.35	.84	.42	.23	.13	.29
8	191.5445	14.42	.80	.36	.91	.45	.21	.11	.25
9	193.6162	17.86	.96	.41	1.12	.65	.25	.13	.32
10	204.3416	25.87	1.07	.46	1.31	.77	.25	.07	.29
Total	208.3183	12.36	.85	.39	.94	.46	.31	.39	.55

2. Summarize and interpret the decile analysis results. Are the patterns in the decile analysis consistent with your conclusions from the logistic regression?

Part III: Lifts and Gains

3. Use the information in Table 1 to create a table showing the lift and cumulative lift for each decile.

4. Create a chart showing the cumulative lift by decile.

5. Use the information in Table 1 to create a table showing the gains and cumulative gains for each decile.

6. Create a chart showing the cumulative gains by decile.

Part IV: Profitability Analysis

Use the following cost information to assess the profitability of using logistic regression to determine which customers should receive a specific offer:

Cost to mail offer to customer:	$.50
Selling price (shipping included):	$18.00
Wholesale price paid by BookBinders:	$9.00
Shipping costs:	$3.00

7. What is the breakeven response rate?

8. What was the gross profit (in dollars, and also as a % of gross sales) and return on marketing for this offer earned by BookBinders from the mailing offering “The Art History of Florence” to all 50,000 customers?

9. Table 3 below summarizes key results for two groups: MAIL = No for those customers whose predicted probability of buying ‘The Art History of Florence’ was less than the breakeven rate, and MAIL = Yes for those customers whose predicted probability is greater than or equal to the breakeven response rate. Included in the table are:
i) the number of customers in each group
ii) the number of buyers of “The Art History of Florence” in each group
iii) the response rate (equal to # buyers divided by # customers) for each group
iv) % of total customers – shows the % of total customers in each group
v) % of total buyers – shows the % of total buyers in each group

What would the gross profit (in dollars, and also as a % of gross sales) and return on marketing have been if BookBinders had mailed the “The Art History of Florence” offer only to customers with a predicted probability of buying greater than the breakeven rate (i.e. those in the MAIL = Yes group)?

Table 3: Summary Statistics by Group (Profitable vs. Not-Profitable to Target)

Bought "Art History of Florence?"

MAIL	# Customers	# Buyers	Response Rate	% of Total Customers	% of Total Buyers
No	34435	1198	.03	68.9%	26.5%
Yes	15565	3324	.21	31.1%	73.5%
Total	50000	4522	.09	100.0%	100.0%