BookBinders: Predicting Response with Logistic Regression

As a direct marketer of specialty books, the BookBinders Book Club has achieved steady growth in their customer base. 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 Dave Lawton, BookBinders' marketing director, to experiment with different database marketing approaches in order improve BookBinders' mailing yields and profits.

Dave 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.

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

Professor Charlotte Mason from the University of North Carolina 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. Professor Florian Zettelmeyer modified this case for use in his course and for use with Stata instead of SPSS. Professor Joonhwi Joo further modified this case to use R in his class.

Copyright 2009 by Charlotte Mason and Florian Zettelmeyer.

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 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.

The company currently has 550,000 customers who are being mailed catalogs. Dave 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." Dave 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

1. Estimate a logistic regression model using "buyer" as the dependent variable and the following as predictor variables:

   last
   *total_*
   gender
   child
   youth
   cook
   do_it
   reference
   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.

   Hint: To do this in R, first transform the buyer and gender variables into a 0/1 dummy variable using data.table syntax

   Then run the logistic regression command.

   Finally, ask R to create a new variable that contains the predicted probability of purchase for each consumer.

2. Summarize and interpret the results (so that a marketing manager can understand them). Which variables are statistically significant? Which seem to be economically important? Interpret the odds-ratios for each of the predictors.

Part II: Decile Analysis of Logistic Regression Results

1. Assign each customer to a decile based on his or her predicted probability of purchase with 'bucket 1' being the highest average purchase probability.
   Hint 1: The "predicted probability of purchase" is the variable "purch_prob" that came out of the logistic regression after you issued the "predict" command. It represents the best prediction of the logit model of how likely a customer is to buy "The Art History of Florence."
   Hint 2: Decile basically means ordered 10 buckets of roughly equal size (1, 2, …, 10). You can rank the customers by their predicted purchase probability using frank, and then use data.table conditional assignment syntax to assign customers into deciles.

2. Create a bar chart plotting the average response rate by decile (as just defined above).

Hint: The "response rate" is not the same as the "predicted probability of purchase." Instead, it is the percentage of customers in a given group (for example a decile) that have bought "The Art History of Florence." In other words, you want "buyer" variable on the y-axis. You will need to use `ggplot` command, combined with `stat_summary()` function to show the average values.

3. Generate a table showing the number of customers, the number of buyers of "The Art History of Florence,' and the response rate to the offer by decile for the random sample (i.e. the 50,000 customers) in the dataset.

Hint: data.table again provides a neat way to calculate summary stats (e.g., sum, mean, sd, …) by groups that we learned in chapter 2. Or, feel free to use “describeBy” of “psych” package that we’ve used before.

4. For the 50,000 customers in the dataset, generate a table showing the average values of the following variables by probability of purchase 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).

5. Summarize and interpret the decile analysis results. Are the patterns in the decile analysis consistent with your conclusions from the logistic regression? (*Hint*: graph some of the results in the previous question.)

Part III: Profitability Analysis

Use the following cost information to assess the profitability of using logistic regression to determine which of the remaining 500,000 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

1. What is the breakeven response rate?

2. For the customers in the dataset, create a new variable (call it "mailto_logit") with a value of 1 if the customer's predicted probability is greater than or equal to the breakeven response rate and 0 otherwise.

Hint: You can use data.table’s conditional assignment syntax

3. Out of the 50,000 test sample, how many customers should have received the targeting promotion mail for “The Art History of Florence” based on the breakeven response rate (i.e., the number of mailto_logit ==1 in the data)? Also, among those who would have targeted, what would have been the response rate (i.e., mean of buyer among mailto_logit ==1)? How much higher is this response rate relative to the overall response rate in the data?

4. Consider that there are 500,000 remaining customers for the roll-out (excluding 50,000 test group in the current data. Assuming our test group is similar to the roll-out group (i.e., our test group is representative of the roll-out group), what is the expected number of buyers of 'The Art History of Florence' if we do targeted mailing based on the breakeven response rate? **Hint:** Count the number of buyers among the targeted (mailto_logit==1) in the test group and multiply it by 10 since the roll-out sample is 10 times larger. Alternatively, you can use the response rate.

Exhibit 1

The BookBinders Book Club
Dataset

Summary information about the BookBinders Book Club's customers' purchasing history and demographics is in the dataset called *bbb.csv.*

Below is a listing of the variable names and descriptions of the data types:

Contents of bbb.csv - contains records for 50,000 customers			
Variable name	Type	Size	Description
acctnum	Numeric	5	Customer account number
gender	String	1	Customer gender - M=male, F=female
state	String	2	State where customer lives (2-character abbreviation)
zip	String	5	ZIP code (5-digit)
zip3	String	3	First 3 digits of ZIP code
first	Numeric	3	Number of months since first purchase
last	Numeric	3	Number of months since most recent purchase
book_	Numeric	8	Total dollars spent on books
nonbook_	Numeric	8	Total dollars spent on non-book products
total_	Numeric	8	Total dollars spent
purch	Numeric	5	Total number of books purchased
child	Numeric	5	Total number of children's books purchased
youth	Numeric	5	Total number of youth books purchased
cook	Numeric	5	Total number of cook books purchased
do_it	Numeric	5	Total number of do-it-yourself books purchased
reference	Numeric	5	Total number of reference books purchased
art	Numeric	5	Total number of art books purchased
geog	Numeric	5	Total number of geography books purchased
buyer	Numeric	1	Did the customer buy "The Art History of Florence?" (1=yes, 0=no)