Kellogg School of Management
Northwestern University

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. 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 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.

Dave'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, Dave 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 Dave 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

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 Professor Florian Zettelmeyer modified this case for use in his course and for use with Stata instead of SPSS..

Copyright © 2015 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 (30 points)

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

   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.

   Hint: To do this in Stata, first transform the gender variable into a 0/1 dummy variable:

   ```
   generate female=(gender=="F")
   ```

   Then run the logistic regression command:

   ```
   logistic buyer last total_ female child youth cook do_it refernce art geog
   ```

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

   ```
   predict purch_prob
   ```

2. Summarize and interpret the results (so that a marketing manager can understand them). Interpret the odds-ratios for **each** of the predictors. Which variables are significant? Which variables are 'important' (at least three most important variables)?

Part II: Decile Analysis of Logistic Regression Results (35 points)

1. Assign each customer to a decile based on his or her predicted probability of purchase. **Hint**: 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." See the "Stata Cheat Sheet" (Working with N-tiles) for help on how to assign

consumers into deciles.

2. Create a bar chart plotting 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." When you graph the “buyer” variable by decile Stata automatically calculates the response rate.

3. Generate a report showing 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.

4. For the 50,000 customers in the dataset run a logistic regression model where you predict response only based on the “child” variable. Why is the odds ratio for “child” different than in the logistic regression in Part I? Please be specific and investigate beyond simply stating the statistical problem.

Part III: Lifts and Gains (20 points)

1. Use the information from the report in II.3 above to create a table showing the lift and cumulative lift for each decile. You may want to use Excel for these calculations.

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

3. Use the information from the report in II.3 above to create a table showing the gains and cumulative gains for each decile. You may want to use Excel for these calculations.

4. Create a chart showing the cumulative gains by decile along with a reference line corresponding to ‘no model’.

Part IV: Profitability Analysis (15 points)

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 “target”) with a value of 1 if the customer’s predicted probability is greater than or equal to the breakeven response rate and 0 otherwise.

3. Considering that there are 500,000 remaining customers, generate a report summarizing the number of targeted customers, the expected number of buyers of ‘The Art History of Florence’ and the expected response rate to the offer by the “target” variable.

4. For the 500,000 remaining customers, what would the expected profit (in dollars) and the expected return on marketing expenditures have been if BookBinders had mailed the offer to buy “The Art History of Florence” only to customers with a predicted probability of buying that was greater than or equal to the breakeven rate?

(Please see the next page for a description of the data)

Exhibit 1

The BookBinders Book Club
Stata Dataset

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

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

Contents of bbb.dta – 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
refernce	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)
training	Numeric	1	Dummy variable that splits the dataset into a training (“1” and validation (“0”) dataset. This variable is used only later in the course.